AN ALTERNATIVE TO THE ORIGINS OF LIFE THEORIES FUNCTIONALLY ENDOWED DNA MOLECULES CAPABLE OF IMPROVED CATALYSIS

BY

DARWIN NOEL ANG

A DISSERTATION PRESENTED TO THE GRADUATE SCHOOL
OF THE UNIVERSITY OF FLORIDA IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

UNIVERSITY OF FLORIDA

1999

ACKNOWLEDGMENTS

I believe that accomplishment, though in large part an individual's effort to persevere, is also credited to those who have influenced one in the direction of life he/she has chosen. I am very fortunate to have many people in my life that have guided me through these enduring years. My success, if any, is a tribute to them.

I would like to begin of course with my family. James, Josephine, and Kevin Ang. I draw strength from knowing I always have them by my side. My parents taught me that the pursuit of my goals should coincide with the pursuit of being a better person. I will always honor my family and conscience first.

Professor Steven A. Benner is my Ph.D. mentor and advisor. The lessons I have learned in his lab will be invaluable to me in my future as a scientist. One day, I hope to be able to be as innovative and creative as Professor Benner. I have also appreciated his generosity and guidance.

Professor Charlotte Hunt was my first mentor and my high school English teacher. Unfortunately she passed away during my graduate studies and will never see me become a doctor like she believed I would. She helped me see my passion in life, To help others. Her suffering due to cancer, diabetes, and congestive heart failure is a constant reminder of why I have chosen a career in both research and medicine.

Professor Michael Kasha was my undergraduate mentor in physical chemistry at Florida State University. He forced me to step up to the intellectual

challenges that he posed to me on a regular basis He has molded the way I think about science and how I think about my own potential in science. I have pursued my graduate degree in chemistry in large part due to him.

Professor Ken Goldsby set the standard for me as being a great research advisor and gave me my first college research experience in his inorganic chemistry laboratory at Florida State University

Dr Joseph Layon is my long time physician mentor and friend who is one of the main reasons why I feel I fit in at UF He has always made sure that I stayed on track

Professor Al Lewis is the director of the M.D./Ph.D program and a member of my committee. Professor Lewis has been extremely supportive as both a research and graduate student advisor Because of his commitment, I feel that Professor Lewis has brought much honor to the M.D./Ph.D program at the University of Florida.

I will always appreciate and remember the guidance of Professor Michial Ross who interviewed me for the M.D./Ph.D program and inspired me to join the Anatomy and Cell Biology Department, Professor Stephen Sugrue who has lived up to all my expectations, and Professors Ben Horenstein and Gerard Shaw who are part of my graduate committee.

Dr Petra Burgstaller was my mentor in *in vitro* selection Dr Jeong Ho Park and Dr Xiao Zhou mentored me in molecular biology techniques.

I am very fortunate to have had Emma Westermann-Clark and Bernard Suh as undergraduate mentees They taught me while I was teaching them. We were a good team. Dr Simeon Jurczyk, Dr Tom Battersby Heike Held and EraGen Inc contributed to the organic chemistry in this project. Mike Thomson was responsible for

the polymerase mutants and help taught the undergraduates many molecular biology techniques (especially cloning). Morgan Peltier taught me to use SAS. Shaojun Zhang offered helpful advice on statistical methods and allowed me to pick his brain on probability theory. I enjoyed both collaborating with and learning from all of them.

Romaine Hughes is the keystone that holds our lab together. Lynda Hanssen made sure I met all my deadlines for registration, paid my tuition fees, and kept up with the graduate program. The American Heart Association Predoctoral Fellowship funded me for a year of my research. The Medical Scientist Training Program at the University of Florida College of Medicine is continuing to fund me through my medical education.

I want to thank the rest of the Benner group (in no particular order) Lee Raley, Daniel Carsco, Mike Sismour, Gina Cannarozzi, Kevin Devine, Rachelle Childress, Thomas Giger, Karen Wilson, Daniel Hutter, Vic Chamma, Stefan Lutz, David Schreiber, Janos Kodra, and Dave Liberles, and the Stewart group (again, in no particular order) Carlos Martinez, Sonia Rodriguez, Jennifer Tonzello, Ann Polyzos, and Kavitha Vedha for all making my studies in Chemistry Building, Leigh 452 a great one to remember.

And last but not least, I thank Hiep and Kirsten Ly who fed me on weekends and for being the closest of friends.

TABLE OF CONTENTS

Page

Abstract of Dissertation Presented to the Graduate School
of the University of Florida in Partial Fulfillment of the
Requirements for the Degree of Doctor of Philosophy

AN ALTERNATIVE TO THE ORIGINS OF LIFE THEORIES
FUNCTIONALLY ENDOWED DNA MOLECULES CAPABLE OF
IMPROVED CATALYSIS

By

Darwin Neal Aug

August 1999

Chairman: Steven A. Benner
Major Department: Anatomy and Cell Biology

This research proposes a novel hypothesis to account for the origins of life. Under this hypothesis, life evolved from DNA molecules endowed with additional chemical groups capable of catalyzing biochemical reactions. Experiments were designed to create catalysts from a random pool of DNA molecules carrying ammonium groups under laboratory-induced selection pressures. The functionalized nucleoside used in these experiments was a lysine like thymidine molecule, 5-(3 amino-1-propynyl)-2 deoxyuridine. This building block contributes a positive charge at neutral pH that is otherwise lacking in DNA. Polymerases were first screened to identify those that could incorporate this nucleoside triphosphate. Mutant polymerases were prepared and studied for their ability to accept the functionally endowed nucleoside triphosphate. *In vitro* selection (SELEX™, Systematic Evolution of Ligands by Exponential Enrichment) experiments were then used to create catalysts from both endowed and non endowed

DNA pools. In both cases, a random pool of 10^{14} oligonucleotides 40 nucleotides in length was exposed to a selection pressure that allowed only a few molecules to survive. The selection pressure favored towards molecules that could cleave a ribonucleotide phosphodiester bond within the backbone of the oligonucleotide. Mutation was introduced in the pool during amplification to promote adaptation to the selection pressure. Overall, eleven rounds of selection were performed.

Catalysts were obtained from both libraries with the endowed library providing faster catalysts sooner in the selection. The intrinsic potentials of both libraries for generating catalysts were determined by fitting the data from the functionalized and non-functionalized DNA pools to a common statistical distribution, the Weibull distribution. Weibull distributions (a) allowed for the prediction and subsequent statistical measurement for the rate of optimization of both pools due to selection pressure, (b) modelled the effect of tightening the selection pressure on a population of catalysts; and (c) quantitated the increase in the catalytic potential of a DNA library attributable to the presence of the ammonium functionality. The calculated intrinsic potential for the amine acid-like DNA pool was better.

Finally, approximately eighty sequences from each SELEX experiment were determined and an evolutionary relationship between the sequences was proposed. Both convergent and divergent evolution was observed in both pools. These findings support a new molecular candidate for a biopolymer that might perform both genetics and catalysis.

BACKGROUND

The origin of life is one of the oldest philosophical questions addressed by humankind. Until recently it has remained a scientific enigma. Philosophy and Religion have been the primary disciplines attempting to appease the intellectual and spiritual curiosity of humanity. More recently, Science has now offered several workable hypotheses based on present biochemical understanding. These began in the first quarter of this century with the Oparin-Haldane hypothesis (Oparin, 1978) which referred to ideas then emerging from the organic chemical analysis of biological systems, and how biochemicals might self aggregate to form structures.

Only after the elucidation of the structure of genetic and catalytic biopolymers was it possible, however, to address the "chicken or the egg" question. This question arises from the fact that contemporary life uses two macromolecules, proteins and nucleic acids, to perform biological function, and that each appears to be required for the synthesis of the other. The probabilities of either proteins or nucleic acids arising spontaneously from a prebiotic soup is low enough. It is inconceivable, however, that both arose spontaneously and simultaneously and each in a form able to assist in the production of the other.

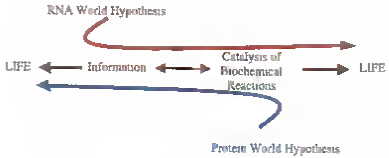
The generic solution to this problem is based on the assumption that one of the two biopolymers was able to survive without the other. This implies that either nucleic acid or protein, alone and without the other, was able to perform both catalytic function and genetic function, presumably assisted by small molecules (including lipids) that

would have emerged spontaneously from abiotic processes. These two extremes are summarized in the form of two hypotheses: The "RNA World Hypothesis" and the "Protein World Hypothesis"

Each hypothesis assumes that life began with its respective biomolecular polymers which eventually evolved into a complex system responsible for self propagation, transfer of information, and evolution. Considering only the biopolymers, the two hypotheses are at opposite ends of the spectrum (Figure 1). The RNA World hypothesis, proposed originally in 1962 by Alexander Rich (Rich, 1962), was supported by the recent discovery of ribozymes, RNA molecules capable of catalyzing biochemical reactions (Gestland et al., 1999). The independent discoveries by Thomas Cech and Sidney Altman that RNA functions as enzymes in biological systems disproved the then prevailing view that oligonucleotides were involved only in information storage. Ribozymes contribute to the pathogenesis of certain parasites (Ferre-D'Amare et al. 1998), and in the processing of genetic information (Green et al., 1998) and even now are being sought as molecular tool to add to our armamentarium in genetic therapy.

The fact that RNA is capable of enzyme-like catalytic activity suggests that an RNA molecule performing both genetic and catalytic functions could have been the precursor of life (Joyce, 1989). Major weaknesses of this hypothesis remain, however. For example, the phosphodiester backbone of RNA molecules is readily prone to hydrolysis due to the 2'-hydroxy group on the ribose ring. Further, RNA molecules are relatively slow catalysts (Bartel and Szostak, 1993).

Figure 1 A Linear model on origins of life the hypothesis proposed by this research lies in the middle of this spectrum.


RNA World Hypothesis
LIFE
Information
Catalysis of
Biochemical
Reactions
LIFE
Protein World Hypothesis

The Protein World hypothesis, on the other hand, assumes that life began with a biopolymer (proteins) already known to provide excellent catalysts. In fact, when we think of biochemical catalysts, protein enzymes are often the first things to come to mind. With the recent discovery by Prusiner of prions (Prusiner, 1997) this hypothesis has been taken seriously as a possible model for the origin of life. Prions may be the cause of a class of diseases known as spongiform encephalitis (Prusiner 1997). In their mechanism of reaction, prions are believed to be proteins that template the folding of other proteins therefore "self-reproducing" their conformation (although not directing their own synthesis).

Recently, Reza Ghadiri's laboratory at the Scripps Research Institute synthesized peptides that could direct their own synthesis (Lee et al., 1956) at least if suitable precursors were provided. These behave as a hypercyclic network (Lee et al., 1997) a collection of two or more molecular systems that are able to aid in the synthesis of each (Eigen, 1971). The self replicating peptides catalyze a template-directed homodimerization of two peptide fragments. Templating relies on the common hydrophobic interaction between certain proteins that leads to a tertiary structure assembled from alpha helices (Lee et al., 1996). The ligation involves the formation of an amide bond through the condensation of an amine a thioester, a simple nucleophile-electrophile reaction.

While intriguing as an example of peptide based replication, Ghadiri's system is still far from "living". First, the chemistry is limited in scope. Further, the two chemical groups (the amine and thioester) involved in the synthesis reaction are very reactive,

would react at a significant rate on their own. However, this novel idea and set of experiments did supply a proof of the concept of protein templating, and spurred a following of believers in the Protein World hypothesis.

In light of the deficiencies in both hypotheses, a third hypothesis was considered, and is examined by the research reported in this dissertation. This hypothesis holds that an early version of life used nucleic acid molecules that were endowed with protein-like functionality. The functionality was covalently attached to nucleosidic building blocks and incorporated into oligonucleotides via template-directed polymerization. An example of a functionalized nucleotide is shown in Figure 2.

This hypothesis combines some of the strengths of both of the two hypotheses outlined above. Specifically, functionalized nucleic acids may have both the heritability and the evolvability evident in RNA molecules. Further, they may have some of the catalytic potential displayed by proteins. If we were to again consider the "chicken or egg" analogy, where for obvious reasons RNA is thought to represent the "egg" and proteins represents the "chicken" this new hypothesis would not only be somewhere in the middle, but in fact, would combine the two together.

Under this hypothesis, the primitive biopolymer carried a broader range of "responsibility" than carried by either proteins or nucleic acids in contemporary life forms. This biopolymer therefore is, at a molecular level, more "complex" than contemporary biopolymers.

Several questions come to mind: What must a complex molecule look like to sustain life? Is it possible for such a molecule to be synthesized under Earth's prebiotic

conditions? Finally, if life actually emerged based on a single biopolymer system, can chemical "fossils" of these complex biopolymers be found in modern metabolism?

These questions correspond to chemical questions. Would the combination of genetic-like structures and reactive amino acid-like functional groups lead to a molecule that, in the laboratory, could display genetic and catalytic properties? Might such molecules arise under non-biological conditions such as those believed to be present on early Earth?

In this work, a 2'-deoxyuridine derivative was prepared with a functional group attached to the 5-position of the heterocycle (see below). It has been recently shown that the synthesis of such derivatives is possible under non-biological conditions (Robertson and Miller 1995 Robertson and Miller, 1995)

Whether these conditions can be viewed as "prebiotic" is open to question. For some time, the prebiotic atmosphere was believed to have been a mixture of CH_4 and N_2 NH_3 and H_2O or possibly CO_2 H_2 and N_2 (reducing atmosphere). In a reducing atmosphere, the synthesis of organic compounds such as amino acids pyrimidines, and sugars (Stribling and Miller, 1987), but also including uridine and formaldehyde two precursors in the Miller-Robertson synthesis is thermodynamically possible. The presence of these and other "biochemical" compounds found on the Murchison meteorite, which is 4.6×10^9 years old, give some possibility that a prebiotic Earth may also have been able to create the same compounds (Anders, 1989)

Atmospheric chemists now believe that the Earth's prebiotic atmosphere was not reducing, but instead was made up largely of CO_2, N_2 H_2O (Lazcano et al., 1996). If this were true the synthesis of organic compounds is highly unlikely to have occurred in the

prebiotic atmosphere. Reducing conditions may have prevailed elsewhere on the primitive Earth, however. Either on or off Earth, the Robertson and Miller experiments provide one model for how functionalized nucleosides might have occurred arisen on early Earth.

Are there signs (or fossil remains) that functionalized nucleotides ever existed? Functionalized nucleotides are actually known in contemporary terrean biological systems. The best examples of this are the more than 80 modified nucleosides characterized from tRNA (Bjork, 1995). Many of these nucleosides are endowed with elaborate functionality. Even though these functionalized nucleosides play an integral role in translation, particularly as "identity determinants" for many aminoacyl tRNA synthetases (Bjork, 1995), their roles *in vivo* are not completely elucidated and it is currently assumed that each type of functionality plays alternative roles.

With results from prebiotic chemistry and the presence of functionalized nucleosides in our own physiology, a functionalized DNA World hypothesis is conceivable. Of course, the true measure of this hypothesis will be whether or not functionalized DNA makes a better candidate for the single biopolymer life form. The research presented here will answer this question in the affirmative.

The method used in this research to create both functionalized and nonfunctionalized DNA catalysts is known as *in vitro* selection or SELEX (Systematic Evolution of Ligands)(Irvine et al., 1991), Figure 3. It is not clear who invented this method first. It is commonly attributed to three pioneering researchers who developed this field of science in the early 1990s, Jack Szostak (Szostak, 1988) at the Massachusetts General Hospital, Gerald Joyce at the Scripps Research Institute (Joyce 1989) and Larry

Gold (Irvine et al., 1991) of the University of Colorado. *In vitro* selection begins with a large pool of random DNA molecules of the same length (e.g. 10^{15} 112mers). The pool is exposed to an experimentally induced selection pressure that singles out specific molecules capable of surviving by virtue of their being able to bind some ligand, catalyze some reaction or perform some other specified task. The molecules that do not survive the selection are eliminated. The few molecules that do survive are collected and amplified by PCR (Polymerase Chain Reaction) methods. The amplified products are then reintroduced into the selection scheme.

Eventually, mutations are introduced via error-prone PCR allowing the DNA molecules to evolve. Mutations that are beneficial or do not affect the selected phenotype of the molecule during selection, survive. Those mutations that decrease property of the molecule demanded by the selection scheme are eliminated (Figure 3).

From this simple method, an impressive array of oligonucleotide enzymes have been selected, particularly novel ribozymes. The first selected RNA molecules were not catalysts. Rather, they were RNA and DNA molecules created as receptors for the purposes of binding to different ligands ranging from dyes to ATP and other biological cofactors (Ellington and Szostak, 1992, Huizenga and Szostak, 1995, Burgstaller and Famulok, 1994, Geiger et al., 1996, Burke and Gold 1997). Although the dissassociation constants (K_d) for these oligomers were relatively high, making them poor receptors, the potential for new receptors spurred excitement in the scientific community. The term "aptamers" is used to describe these RNA receptors (Ellington and Szostak, 1992).

Figure 2: 5-(3-amino-1propynyl)-2 -deoxyuridine or J base

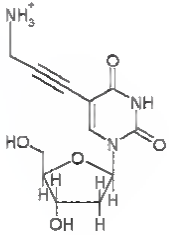
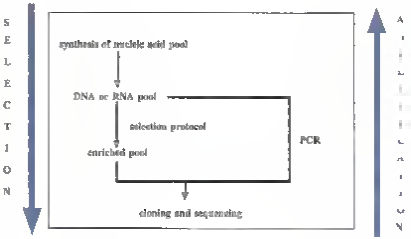

NH3+
O
NH
N
O
HO
O
H
H
H
H
OH
H

Figure 3 A general outline for all *in vitro* selection experiments.


SELECTION
synthesis of nucleic acid pool
DNA or RNA pool
selection protocol
enriched pool
PCR
cloning and sequencing
AMPLIFICATION

The evolution in the science of aptamers has taken on several forms. Researchers began by seeing *in vitro* selection as a way of making receptors to biological molecules. Aptamers that bound to thrombin, at least *in vitro*, decreased the rate of formation of fibrin clot (Bock et al. 1992). Several other aptamers bound to elastase (Smith et al. 1995) these proved to be modest inhibitors of elastase activity. Currently, these inhibitors are being tested *in vivo*.

The need to have stable aptamers for application in mammalian systems led to the next stage in the evolution of aptamers research. RNA molecules are relatively unstable. RNA will hydrolyze spontaneously. Further, ribonucleases are ubiquitous in any *in vivo* environment. For both reasons, RNA aptamers provided exogenously have difficulty finding a serious role as therapeutic agents. To circumvent this problem researchers created the mirror image of RNA oligomers. These enantiomers were found to be resistant to RNA nucleases (Nolte et al. 1996).

Ribozymes created via *in vitro* selection methods did not attract much attention until David Bartel and Jack Szostak (Bartel and Szostak 1993) created the first self-ligating ribozyme. This landmark experiment showed how powerful *in vitro* selection could be as a tool for creating actual enzyme catalysts. Chemists and biologists alike began to ask how far the biology and chemistry of an oligonucleotide might go when subjected to natural selection and evolution (Joyce, 1997). The implication that ribozymes could be intentionally selected in the laboratory to self-ligate led to the search for broader catalysis. In particular, it was asked whether self polymerization might be possible and, more specifically, a self-replicating nucleic acid (replicase)? Is the scope of

RNA catalysis able to go outside of phosphoester chemistry? And what types of molecules (based on function) need to be synthesized in order to provide proof of concept for the RNA World hypothesis?

Since then, *in vitro* selection has created ribozymes capable of self-alkylation (Wilson and Szostak, 1995), polymerization of triphosphates (Ekland and Bartel, 1996) self aminoacylating RNA (Illangasekare et al., 1997) synthesis of peptide bonds (Zhang and Cech 1997) synthesis of nucleotides (Unrau and Bartel, 1998), and two examples of ribozymes capable of synthesizing carbon-carbon bonds via Diels/Alders reaction (Tarasow et al., 1998 Seelig and Jaschke, 1999)

It is evident that the range of *in vitro* selected catalysts and splinters has been broad Much less effort has been devoted to asking how nucleic acid-based catalysts function, and whether *in vitro* selection as a tool might be improved. Only very recently have researchers began to determine at a molecular level (or a quasi-atomic level) what reaction mechanisms may be followed in RNA-based catalysis These have exploited microscopic rate constants and substrate modification, for example as tools (Santoro and Joyce, 1998, Esteban et al., 1998, Lott et al., 1998) Generally, it appears that the divalent cation present in the majority of these nucleozenzymes is involved in catalysis In other cases metal ions are known to play a structural role (Hampel and Cowan, 1997) Some nucleozenzymes apparently have no requirement for metal (Geyer and Sen, 1997) Nevertheless an overwhelming number of these nucleozenzymes do require a metal cofactor for catalysis.

The first work has now appeared to describing the crystal structures of naturally occurring ribozymes such as the Group I ribozyme domain (Cate et al., 1996) the

hepatitis delta virus ribozyme (Ferre-D'Amare et al., 1998), the tetrahymena ribozyme (Golden et al., 1998), and the hammerhead ribozyme (Scott et al., 1995; Scott et al., 1996) (Doudna, 1997). Thus offers a new, not yet fully exploited, approach to understand the catalytic mechanisms of nucleoenzymes.

From an analytical or mathematical perspective, even less literature can be found quantitating the power of *in vitro* selection. Ultimately, we want to know how frequently a catalyst of a particular power appears in a oligonucleotide population. This is not a straightforward question to answer. It is impossible to search all of the possible sequence combinations for even modest sized oligomers, say 100mers, which number approximately 1.6×10^{60} molecules. To put this number into perspective, the Universe has only has only ca. 10^{43} hydrogen molecules in the Milky Way Galaxy (Britannica, 1986). Obviously, only a small fraction of the possible sequences catalysts can be sampled.

From Bartel's landmark ligase ribozyme selection, one could say that the number of ribozymes he discovered compared to the number of sequences he sampled would give a quantitative statement that says, a pool of 10^{14} RNA 220mers must be sampled before one catalyst can be found that is capable of ligating itself. For RNA-cleaving DNA molecules, one only needs to search a pool of 10^{12} 100mers before finding one molecule capable of hydrolyzing an RNA molecule (Breaker and Joyce, 1995). At first glance, it seems obvious that because the ratio of number of catalysts found to number of oligonucleotides in the pool is much higher for RNA cleaving molecules, that RNA hydrolysis can be found more easier than RNA ligation. But is this really true? In sequence space, Breaker sampled 1×10^{13} molecules out of a possible 1×10^{60}, while

Bartel only searched 1×10^{14} out of a possible 1×10^{120}. In fact, if we compare these ratios, we find that the search done by Breaker in sequence space is several tens of order of magnitude (10^{45}) better than Bartel's. Of course this is complicated by the fact that the *in vitro* selection community believes that ligation is a more difficult reaction and therefore the requires a longer oligomer. In reality, there should be no significant difference since the ligation of RNA's phosphodiester bond is, chemically, the microscopic reverse reaction of the hydrolysis of RNA's phosphodiester bond. In principle, one should be able to find a ligase just as easily from a pool of 10^{13} 100mers, as they will find an RNA phosphodiesterase.

In a review article by Ronald Breaker, he introduces the concept of how many contiguous sequences can be represented in a stretch of oligonucleotides with the mathematical transformation, $(x - y) + 1$, where x is the number of nucleotides in the random region and y is the number of nucleotides of interest (Breaker 1997). With this simple equation one can determine how long an oligomer needs to be in order to effectively represent, not only the length of the oligomer of interest, but how many need to be synthesized or used for the selection.

The goal of this dissertation was to develop the tools needed to do *in vitro* selection with functionalized nucleotides, to apply these to generate a functionalized DNA catalyst, and to develop the quantitative tools to assess how much the addition of protein-like functionality improved the catalytic potential of nucleic acids. The dissertation work began with the screening of polymerases, the from the biochemical data supporting an alternative single biopolymer life form, to mathematical models describing the improvement of functionalized DNA over nonfunctionalized DNA. This is due to the

fact that a novel field within science is being proposed whereby a multidisciplinary approach is necessary. Structural biology, probability theory and the biochemistry of DNA are among the few topics addressed in this research. But it is worth noting that however diverse the topics are, it is the aim of this project to focus these areas to answer the question: Is it possible to improve DNA?

The functionalization introduced into the *in vitro* selection experiments reported here is 5 (3 amino-1 propynyl)-2 -deoxyuridine (Figure 2). This molecule carries a lysine-like amino group attached to a 2 -deoxyuridine nucleotide and was given the trivial name dU. At physiological pH the amino group is protonated, and therefore carries a positive charge. This functional group was chosen because natural nucleotides lack chemical groups carrying a positive charge. With the introduction of a positive charged amino group DNA's electrostatic properties are expanded, and should make DNA a more versatile molecule.

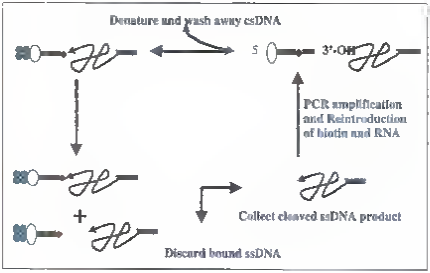
The selection scheme for this project was designed to create a catalyst that cleaves the phosphodiester bond of a RNA. To control the position where cleavage occurs simplifying the analysis of products, the substrate for the reaction was a DNA molecule 51 nucleotides in length containing a single ribonucleotide. The ribonucleotide was embedded within the 5 end of a primer region of the DNA molecule.

The library consisted of 10^{15} molecules 112 nucleotides long. The molecules in the library were designed for the purposes of PCR amplification. Each molecule possessed a 40 nucleotide long random region flanked by two defined primer binding sites. At the 5 -end of the strand of interest a biotin molecule was appended so that the DNA strand could be immobilized on a streptavidin column.

Molecules that were able to catalyze the cleavage of a RNA DNA phosphodiester bond could release themselves from the biotin tag, and therefore from the streptavidin column. These molecules were said to have "survived" the selection and were collected and reamplified. Molecules that could not catalyze the cleavage remained immobilized on the support and subsequently discarded.

The biotinylated primer and ribonucleotide were reintroduced during PCR of the survivors, and the molecules were repeatedly exposed to the selection. Figure 4 illustrates the selection scheme for this experiment.

Figure 4 The selection scheme for the cleavage of a ribonucleotidyl bond The green circles represent the streptavidin column the oval represents the biotin molecule at the 5'end of the primer binding site of the oligonucleotides, and the red diamond represents the single ribonucleotide embedded within the primer binding site. The oligonucleotide pool is immobilized to the support Molecules that catalyze the cleavage are collected and amplified by PCR, while those that are still immobilized on the column are discarded. The PCR reintroduces the biotin moiety and the ribonucleotide to the cleaved DNA The double stranded DNA is immobilized to the support the complementary strand (csDNA) is denatured and washed away The single stranded (ssDNA) is reintroduced to the selection pressure


Denature and wash away csDNA
5
3'-OH
PCR amplification and Reintroduction of biotin and RNA
+
Collect cleaved ssDNA product
Discard bound ssDNA

The first step towards performing *in vitro* selection using functionalized nucleotides required the identification of a polymerase that would incorporate dU via template directed polymerization specially during PCR reactions The initial method used in this study employed random screening of commercial polymerases. These were chosen from both Family A and Family B polymerases two evolutionary families of polymerases family A and B defined by a phylogeny created by Braithewhite and Ito (Braithewhite and Ito 1991) This classification is based on the alignment of protein sequence data. The families A B and C correspond to DNA polymerase I II and III

For those polymerases that showed promise, kinetic studies were done to quantitate the impact of the functionality on the ability of the polymerase to accept a nucleotide triphosphate These polymerases were also probed using a modified form of the J-base wherein the amino group was protected as its phenylacetyl amide This protected side chain lacked the positive charge of the amine

A series of molecular modelling experiments were done by Bernard Salt to build hypotheses concerning contacts between specific amino acids in the polymerase active site the functional group appended to the nucleotide The sequence of a family B polymerase (used for the kinetic study, Vent) was threaded on to the crystal structure of a highly homologous (> 80% identical), but not commercially available family B polymerase (*Thermococcus gorgonarius*) The geometry was subjected to a local energy minimization.

Finally in an effort to confirm hypotheses based on molecular modelling, key amino acid residues were mutated by J Michael Thomson on both family B and family A polymerases and kinetic analyses were performed

EXPERIMENTAL DATA AND RESULTS

Developing a New Methodology and An Evolutionary Prediliection of Polymerases that can Use an Protein like Deoxynucleotide

This research posed a unique challenge since it is the first time modified nucleotides would be employed to create deoxyribonucleotide catalysts. As a result, a method had to be developed to incorporate modified nucleotides by polymerases. Identifying polymerases that incorporated the dU-base via template directed polymerization and assessing the kinetic behavior of these polymerases was an important first, as well as the rate limiting, step for those experiments.

Primer Extension Screening and PCR Amplification. An Evolutionary Predilection Resolved.

Many polymerases were screened and several polymerases were found to incorporate the dU base during primer extension experiments. Immediately an interesting observance was made. There appeared to be a correlation between the ability of a polymerase to incorporate the dU base, and its evolutionary pedigree. All Family B polymerases screened were able to use the dU base as a substrate and create full-length product. In contrast, most Family A polymerases produced truncated products (Figure 5).

Thermophilic polymerases were screened by PCR to determine whether or not they could repeatedly use the dU-base as a substrate for the purpose of amplification. From these experiments, it was shown that Vent was the best polymerase from the several thermophiles screened. dUTP was incorporated in a PCR experiment by Vent polymerase only ca. 70% as efficiently as TTP (Figure 6b), as judged by a comparison of the rate of increase in the intensities of full length product bands on an ethidium bromide stained agarose gel (Figure 6a). Polymerase chain reactions that lacked either the TTP or dUTP nucleotide were run in parallel and served as negative controls. In both cases, no full-length product was seen after eight cycles of PCR amplification.

Determining the Kinetic Behavior and Substrate Modification

The kinetic behavior of Vent exo- polymerase accepting dUTP as a substrate was then determined. All kinetic experiments were done with a running-start primer-template design and performed as described by Creighton and Goodman (Creighton, 1995). The extended primer and its products were quantitated by phosphorimaging the ^{32}P labeled primer.

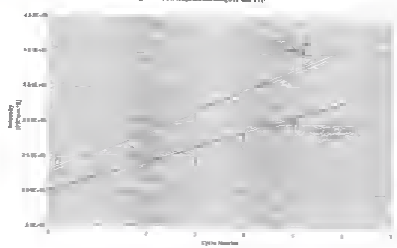
The amount of the full length product was interpreted using the equation $I^S_T/I_T = V_{max}*[dNTP]/K_m+[dNTP]$, where V_{max} is the relative maximum velocity for incorporation of TTP or dUTP at the target site A. The kinetic data were plotted assuming Michaelis-Menten behavior (Figure 7). The results are collected in Table I.

Figure 5 Polymerase B and A families screened against TTP and dUTP by primer extention experiments The family B polymerases (lanes 1 10) show full length product whilemany of the family A polymerases (lanes 11 20) show truncation. Lanes 1 40 are family B polymerases lanes 11 20 are family A polymerases lanes 21 25 are the TTP products of family B polymerases lanes 26-30 are the dUTP products of family A polymerases: lane1 is Vent exo- w/dUTP lane 2 is Vent exo- w/TTP lane 3 is Deep Vent exo- w/dUTP lane 4 is Deep Vent w/TTP lane 5 is T4 w/ dUTP lane 6 is T4 w/ TTP lane 7 is PFU w/ dUTP, lane 8 is PFU w/ TTP lane 9 is PWO w/ dTP lane 10 is PWO w/ TTP lane 11 is BST w/ dUTP lane 12 is BST w/ TTP lane 13 is Pol I w/ dUTP lane 14 is Pol I w/ TTP lane 15 is Taq w/ dUTP lane 16 is Taq w/ TTP lane 17 is Tth w/ dUTP lane 18 is Tth w/ TTP lane 19 is T7 w/ dUTP lane 20 is T7 w/ TTP lane 21 is Vent exo- w/ JTP lane 22 is Deep Vent w/ JTP lane 23 is T4 w/ JTP lane 24 is PFU w/ dJTP lane 25 is PWO w/ dJTP, lane 26 is BST w/ dJTP lane 27 is Pol I w/ dJTP lane 28 is Taq w/ dJTP, lane 29 is Tth w/ dJTP, lane 30 is T7 w/ dJTP


Primer
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30

Figure 6 (A) Eight separate cycles of PCR amplified product shown on 4% agarose gels. The two end lanes on either side of the cycles are 50 base pair ladders. Both gels were stained with etidium bromide. (B) The comparison of the slope of the intensities between the dU base (functionalized pool) versus the standard (nonfunctionalized pool). The red data points represent the functionalized pool and the blue represents the nonfunctionalized pool. The ordinate is intensity of amplification as a function of PCR cycle (abscissa).


A
J-base
Cycle 0 1 2 3 4 5 6 7 8
TTP


B

The relative apparent V_{max} of the polymerase with dUTP as substrate was 70% that of apparent V_{max} for TTP. To determine whether the difference in apparent V_{max} values could be attributed to either the presence of the positive charge from the amino group or the mere presence of the side chain, analogous kinetic experiments were performed with a dUTP derivative where the amino group was protected as the uncharged phenylacetyl amide. The apparent V_{max} of the protected derivative was the same as that for TTP, the apparent K_M of the protected derivative was approximately 2.5 fold higher than that for TTP restored. Table 1 shows the kinetic values for Vent exo- and the three separate substrates; TTP, dUTP, and protected dUTP (Copeland, 1995)

This result was surprising, as the side chain of the protected form of dUTP is bulkier than the side chain of dUTP itself. These results suggest that charge, not bulk, accounts for the diminished potential of dUTP to serve as a substrate for Vent polymerase.

A Proposed Mechanism and Crystal Structure for the Vent Polymerase and J-base Interaction

The evident correlation between the ability of a polymerase to accept dUTP and its evolutionary pedigree prompted a closer examination of the crystal structure of the polymerase active site (Wang et al, 1997). All family B polymerases have three aspartate residues in the active site. These have been theorized to be responsible in part for catalysis of the formation of phosphodiester bond (Blanco, 1995, Wang et al., 1997). Two of these are in Region I, one is in Region II. Most of the Family A polymerases, however, have only one aspartate in Region I and one in Region II.

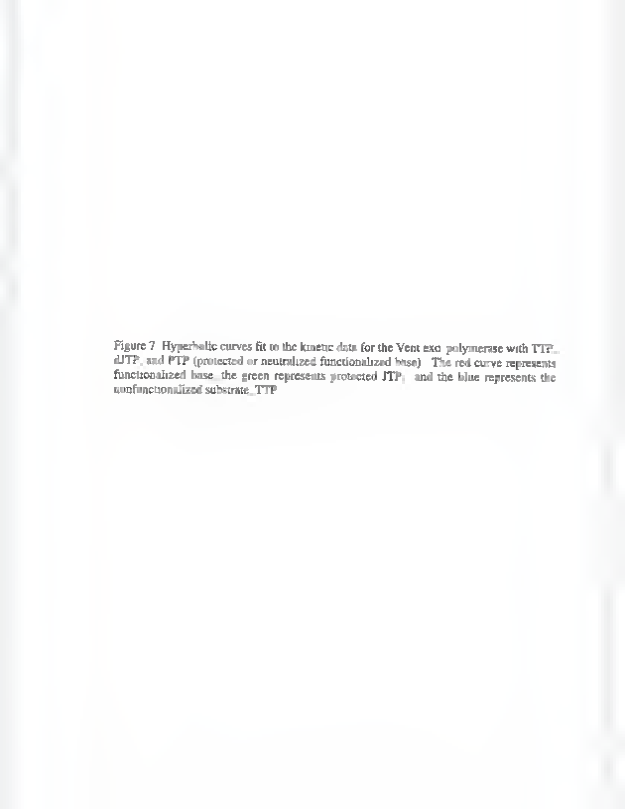


Figure 7. Hyperbolic curves fit to the kinetic data for the Vent exo- polymerase with TTP, dTTP, and PTP (protected or neutralized functionalized base). The red curve represents functionalized base, the green represents protected PTP, and the blue represents the nonfunctionalized substrate, TTP.

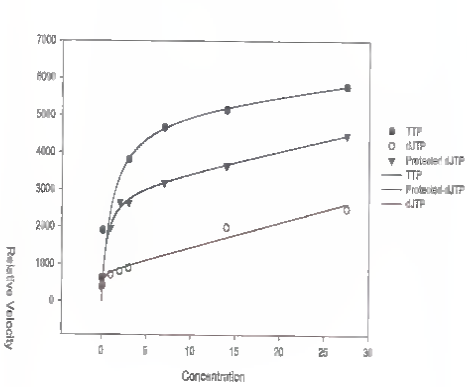

Michaelis-Menten for Vent exo Polymerase
7000
6000
5000
4000
3000
2000
1000
0
Relative Velocity
0
5
10
15
20
25
30
Concentration
TTP
dUTP
Protected dUTP
TTP
Protected dUTP
dUTP

Table 1 Apparent kinetic values for running start experiment with Vent exo-

	Km (M)	Std Dev Km	Rel Vmax	Std Dev Vmax
TTP	4 026	1.298	1 845	0 188
di-base	8 336	2 618	0 612	0 082
Pr- Base	10 088	2 75	1 897	0 229

An alignment of active site of the polymerases used in this project and other polymerases of interest are shown in Figure 8. The first 12 polymerases are from Family B polymerases; the last 11 are from Family A. The sequences were obtained from GenBank and an initial alignment was performed by CLUSTAL W. Because of low sequence similarity between the families, the alignment in the reference Wang et al. (1997) was used to guide the alignment used here.

Inspection of the crystal structure suggested the following model to explain the ability of Family B polymerases to accept dUTP, and the inability of Family A polymerases to do so. In this model, the third aspartate in Region I of the Family B polymerases is expendable for catalysis. It can therefore interact electrostatically with the positively charged amino functional group of dUTP. This interaction occupies the attentions of the positive charge of the amino group, which (so occupied) does not form a coulombic interactive with one of the other two aspartate residues, both of which are essential for catalysis (Steitz, 1998). In contrast, in Family A polymerases, there is no expendable aspartate to interact coulombically with the ammonium group of dUTP. Therefore, the ammonium group of dUTP interacts with one of the critical aspartates, inhibiting catalysis.

The simple prediction of this model is that Family A polymerases would accept protected dUTP. A recent report by Barbas and his coworkers shows that this is the case for a similar substrate 3-pyridynyl 2-enone deoxyuridine triphosphate (Sakthivel and Barbas, 1998).

The two catalytically important aspartate residues in Vent polymerase are numbered 407 and 545. In theory, the metal ion conjugated to aspartate residue 407 would lower the affinity of the 3'-OH group of the primer, allowing the oxygen to perform a nucleophilic attack on the alpha phosphate of the triphosphate. The metal ion coordinated to aspartate residue 545 would facilitate the pyrophosphate in leaving. According to Steitz, the two metal ions stabilize the structure and charge of the expected pentacovalent transition state during catalysis.

A computer generated model (created by using Sybyl) of the crystal structure of the J-base interaction in the active site of the Vent polymerase was created by threading the Vent polymerase sequence to a highly homologous and recently crystallized family B polymerase, the *Thermococcus gorgonarius* (Tgo) (Hopfner, 1999), see Figure 9. The Tgo polymerase has over an 80% sequence identity to the Vent polymerase and its active site for catalysis is identical. Sequence alignment and identification of the active site was determined by previous literature (Braithwaite and Ito, 1991, Wang et al., 1997) and by CLUSTAL W (Thompson 1994).

As indicated by the proposed structure, the functional group of the J-base is in closest proximity to the Asp 543. An electrostatic interaction is highly plausible. The functional group is also very close to the other two aspartate residues thought to be the responsible for catalysis, or specifically, the coordination of the metal ion. It is visually conceivable from the crystal structure that without the aspartate residue 543 to "draw or pull" the functional group away from the other two aspartate residues, an unwanted interaction between the positively charged amine and negatively charged aspartate residues responsible for coordinating to the metal ions may occur.

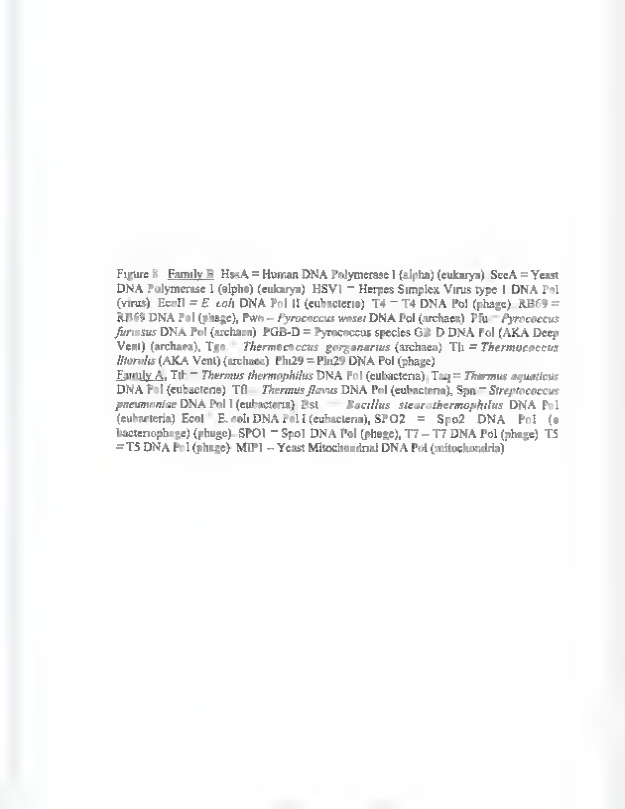


Figure 8 Family B HsaA = Human DNA Polymerase I (alpha) (eukarya) SceA = Yeast DNA Polymerase 1 (alpha) (eukarya) HSV1 – Herpes Simplex Virus type 1 DNA Pol (virus) EcoII = *E. coli* DNA Pol II (eubacteria) T4 – T4 DNA Pol (phage) RB69 = RB69 DNA Pol (phage), Pwo – *Pyrococcus woesei* DNA Pol (archaea) Pfu – *Pyrococcus furiosus* DNA Pol (archaea) PGB-D = Pyrococcus species GB D DNA Pol (AKA Deep Vent) (archaea), Tgo = *Thermococcus gorgonarius* (archaea) Tli = *Thermococcus litoralis* (AKA Vent) (archaea) Phi29 = Phi29 DNA Pol (phage)

Family A, Taq – *Thermus thermophilus* DNA Pol (eubacteria), Taq = *Thermus aquaticus* DNA Pol (eubacteria) Tfl – *Thermus flavus* DNA Pol (eubacteria), Spn – *Streptococcus pneumoniae* DNA Pol I (eubacteria) Bst – *Bacillus stearothermophilus* DNA Pol (eubacteria) EcoI – E. coli DNA Pol I (eubacteria), SPO2 = Spo2 DNA Pol (a bacteriophage) (phage). SPO1 – Spo1 DNA Pol (phage), T7 – T7 DNA Pol (phage) T5 = T5 DNA Pol (phage) MIP1 – Yeast Mitochondrial DNA Pol (mitochondria)

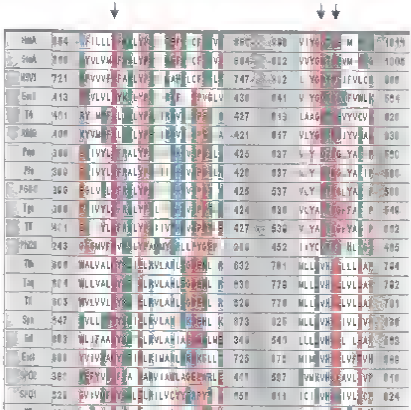

Region II
Region I

Figure 9 The family B active site of catalysis (or catalytic pocket) showing the J base functional group (purple) in close proximity to the aspartate residues (yellow). The aspartate residue closest in proximity to the J base amino propynyl arm is aspartate 543. The green molecules are primer template and the white lines represent the general infrastructure of the Vent polymerase active site.

Polymerase Mutagenesis and Functional Analysis

To test the hypothesis that the "third" Asp residue is important, if not essential for the incorporation of dITP, mutations in the active site of Taq polymerase were introduced. *Taq (Thermus aquaticus)* was chosen as a representative thermophilic polymerase from Family A. It was an ideal candidate because it did not incorporate the *I* base during primer extension screening and a crystal structure of Taq existed (Kim et al., 1995).

In the first mutant, Taq (V583D) an aspartate was introduced at the position analogous to the position holding the third aspartate in Vent. The second Taq mutant, Taq (V583D E786G) was a corresponding double mutation where Val 583 was replaced by an aspartate, and Glu 786 by a glycine.

Effectively, the double mutations placed in the active site of Taq the conserved residues from the Vent active site. If the model outlined above is correct, then the Taq polymerase should be able to accept dITP. In the single mutant variant of Taq, an extra negatively charged residue is present in the active site. There is no prior literature to generate expectations regarding the outcome of this experiment.

Running start experiments were done using both the Taq (V583D) and Taq (V583D E786G) and the primer template combination shown in Figures 10 and 11 over a range of typical triphosphate concentrations. The Taq (V583D) single mutant was not able to incorporate either dITP (beyond a small amount) or TTP. The Taq (V583D E786G) variant was however able to incorporate TTP with an apparent Vmax only ca. 2 fold less than that for TTP displayed by the wild type Taq polymerase.

To explain these results, a simple model notes that the Taq (V583D) variant has two anionic groups in the active site binding pocket, the one there naturally (Gly 786) and the one introduced artificially. Naively, one might argue that placing an anion at either position 583 or position 786 generates an active enzyme, but placing anionic groups at both destroys activity possibly through coulombic repulsion.

The model cannot be so simple, however, as the Taq (V583D) polymerase had sufficient catalytic activity that it could incorporate two guanosines in the running start experiment. Thus, the two negative charges in the the same active site evidently do not prevent the enzyme from incorporating dGTP.

Interestingly, the double mutant did show a trace amount of incorporation of dTTP at high concentrations. This trace was not observed with wild type Taq. While these results require further examination, this is a finding that deserves further exploration.

Figure 10 Primer extension at various concentrations of target nucleotide TTP and dI for purified Taq wild type and primer extension at various concentrations of dI for Taq single and double mutants The running start nucleotide, dGTP is incorporated twice before the target nucleotide dTTP (A) wild type Taq lane 1 10 μM TTP; lane 2 60 μM TTP lane 3 100 μM TTP; lane 4 10 μM dITP; lane 5 60 μM dITP lane 6 100 μM dITP (B) Taq single mutant lane 1 primer; lane 2 0-1 μM dITP; lane 3 1 μM dITP; lane 4 2 μM dITP lane 5 3 μM dITP lane 6 7 μM dITP; lane 7 14 μM dITP lane 8 25 μM dITP lane 9 40 μM dITP lane 10 60 μM dITP; lane 11 100 μM dITP (C) Taq double mutant lane 1 primer; lane 2 0 1 μM dITP lane 3 1 μM dITP lane 4 2 μM dITP; lane 5 3 μM dITP; lane 6 7 μM dITP; lane 7 14 μM dITP lane 8 25 μM dITP; lane 9 40 μM dITP lane 10 60 μM dITP; lane 11 100 μM dITP

A. Taq wild type TTP and dJTP

B. Taq single mutant (V583D) with dJTP

C. Taq double mutant (V583D E786G) with dJTP

Figure 11 Primer extension at various concentrations of target nucleotide TTP for for Taq single and double mutants The running start nucleotide, dGTP is incorporated twice before the target nucleotide dTTP (A) Taq single mutant lane 1 primer lane 2 0 1 μM TTP lane 3 1 μM TTP lane 4 2 μM TTP lane 5 3 μM TTP lane 6 7 μM TTP lane 7 14 μM TTP lane 8 25 μM TTP lane 9 40 μM TTP lane 10 60 μM TTP lane 11 100 μM TTP (B) Taq double mutant lane 1 primer lane 2 0 1 μM TTP lane 3 1 μM TTP lane 4 3 μM TTP lane 5 14 μM TTP lane 6 25 μM TTP lanes 7 40 μM TTP lane 8 60 μM TTP

D. Taq single mutant (V583D) TTP

E. Taq double mutant (V583D E786G) TTP

***In Vitro* Selection of Phosphodiester Cleavage by DNA**

Parallel SELEX™ experiments were performed using pools of modified and non-modified DNA oligomers. In both experiments, novel catalytic DNA molecules were obtained and characterized. The selection was performed as described in the *Background*, and followed as closely as possible the procedure reported in the Breaker and Joyce experiments (Breaker and Joyce 1997).

Both pools were radiolabeled by incorporating $dAT^{32}P$ during PCR amplification. The pools were immobilized on streptavidin-agarose columns, which were repeatedly washed with Tris EDTA buffer at pH 7.6 to rid the columns of nonbinding DNA molecules. The column was treated with NaOH to disrupt duplex DNA structures and immediately washed several times with reaction buffer without divalent cation (pH 7.6). The repeated washings with the neutralization buffer assured that the column was free of unincorporated DNA and radiolabeled dATP. The single stranded DNA immobilized on the column was then incubated with reaction buffer for periods ranging from 15 minutes to 2 hours. The eluted single stranded DNA was collected and ethanol precipitated. Parallel test PCRs were performed to determine how many cycles were necessary for each pool to generate a single band of equal intensity on agarose gel. Once the number of cycles was established, large scale preparative PCRs were carried out, half of the product was used for the next round of selection, and half was used for storage and sequencing reactions.

Catalytic activity was recorded as the amount of radioactivity, or counts per minute, collected from the column after the selection protocol. Since DNA molecules having no catalytic activity remained on the support, the radioactivity eluting from the

column represented those DNA molecules that had catalytic activity. Both the standard and dU-containing pools were PCR amplified with alpha labeled $dAT^{32}P$.

The progress of the selection was followed by determining the amount of radioactivity eluting under the selection conditions after each round. In total, 11 rounds of selection were performed. The fraction of the library released by self-catalysis from the support after each selection interval began to increase detectably after round 5 (Figure 12). This increase was greater for the functionalized pool than for the nonfunctionalized pool in all rounds of selection (with significant DNA elution). Of the 11 rounds of selection, 7 rounds were done under conditions favoring random mutation via error prone PCR. Error prone PCR was achieved by using Vent polymerase without proofreading function, or Vent exo-. The mutation rate was approximately 0.02% per nucleotide per round of selection after fifteen rounds of PCR amplification.

Radiolabeled PAGE (polyacrylamide gel electrophoresis) showed that both the dU containing pool and the nonmodified pool were able to perform intramolecular catalysis. A ^{32}P labeled primer was used as substrate, the substrate was incubated with DNA pools for 2 hours under conditions identical to those used for the selection, and the appearance of cleaved product was followed by PAGE (Figure 13). The multiple bands seen in the gel for the cleaved product are due to the heterogeneous length of the substrate. Subsequent analysis with a radiolabeled ladder indicated that the site of cleavage is at or near the ribonucleotide phosphodiester bond.

Independent amplification of the pool derived from the dU-containing library but using TTP instead of dUTP, gave a mixture that lacked catalytic activity. This indicated that catalysis in the dU-containing pool required the amino acid-like functionality.

Figure 12 Percent cleaved or eluted after eleven rounds of selection The blue graph represents the unfunctionalized pool the red graph represents the functionalized pool

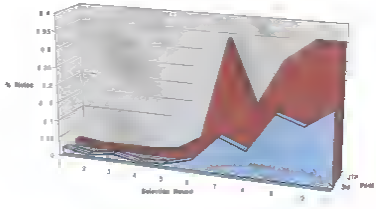
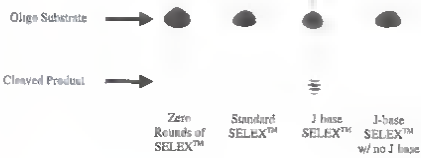
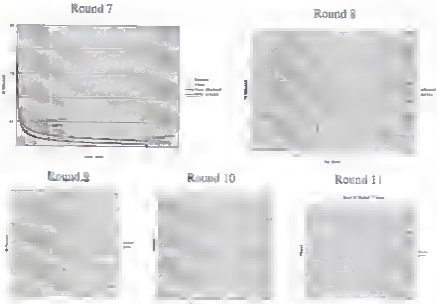

% Noise
Selection Round
Pred

Figure 13 Radiolabeled PAGE showing catalysis after seven


Oligo Substrate
Cleaved Product
Zero Rounds of SELEX™
Standard SELEX™
J base SELEX™
J-base SELEX™ w/ no J base

Experiments were done with both pool DNA and individual sequences to test the role of magnesium cation in the reaction. No detectable cleaved product was observed after 24 hours when the reaction buffer was devoid of divalent magnesium cation. The mechanism for hydrolysis of the phosphodiester bond may involve divalent cation as either a general base or Lewis acid. As previously hypothesized (Santoro, 1998) for these RNA cleaving deoxyribozymes, the bases themselves rarely participate in the actual catalysis. They provide instead the infrastructure for the divalent cation to perform the catalysis. The presence of more cleaved product for the dI-containing pool in the radiolabeled gel led to preliminary indications that the dI-containing pool contained a higher percentage of faster catalysts. This assumption was later empirically confirmed by comparing the data between functionalized and nonfunctionalized pools from selection rounds 7 through 11 (Figure 14). Figure 14 shows percentage of pool cleaved versus rate (min^{-1}). After every selection round, the dI-containing pool generates more product than the standard base pool. These experiments were repeated and the behavior of these curves are mathematically analyzed in later sections of this chapter.

Figure 14 *In vitro* selection rounds 7-11. The red data points represent the functionalized pool and the blue represents the nonfunctionalized pool. In each graph, the Y-axis is the percent cleaved and the X-axis is rate (it increases to the right)


Round 7
Round 8
Round 9
Round 10
Round 11

Analysis of Sequences from Both the dJ-containing and Standard Base Pools

Sequence Analysis

Sequences have been determined for each pool (functionalized and nonfunctionalized) and an evolutionary relationship was derived by Emma Westerman Clark for both the dJ-containing and the non-modified SELEX experiments (Figure 15). Remarkably, both convergent and divergent evolution are observed. Specifically, molecules that are clearly non homologous to each other converged to the same phenotype, or perform the same catalytic function. Molecules that have homologous genotypes but vary in phenotypic character or perform catalysis to varying degrees give evidence to divergent evolution. The sequences between the dJ containing pool and the standard base pool did not show any significant sequence homology. When substituting the dJ-containing for thymidine in the standard base pool, an noticeable cleavage reaction occurred. Both the standard and J base pools were chemically synthesized to begin the selection with an average nucleotide rate of 1:1:1:1. These oligonucleotide pools were purchased and made special order by Integrated DNA Technologies. The distribution of base ratios after seven rounds of selection were determined and given in Table 2.

The nucleotide distribution for the nonfunctionalized selection showed that all four nucleotides had a ratio 1:1:1:1. The functionalized nucleotide distribution showed a "G" rich predilection and a "C" poor construction. For the functionalized pool the "A" and "dJ" bases did not fall outside of one standard deviation from a 25% composition of the total nucleotide population. These findings suggest that the secondary and tertiary structures derived from guanosine nucleotides may play an important role in catalysis for the dJ-base library. It is cited repeatedly in literature that guanosine is abundant in DNA

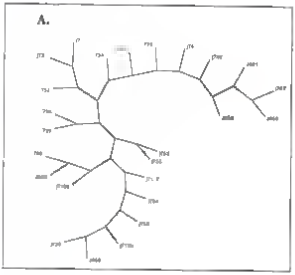
aptamers (oligonucleotides selected for binding) (Huizenga 1995, Sassanfar 1993). However, the guanosine rich phenomenon is not always the case for deoxyribozymes. Since the amino acid functional group is also essential for catalysis it could be possible that adding functionality also confers a greater propensity for more complex tertiary structures.

Kinetic analysis of specific sequences for J base and nonmodified pools

Sequences were chemically synthesized and tested for individual rates of catalysis. Overall seven sequences were sampled from each pool. Four sequences from the standard base pool had half lives recordable within a twenty four hour time interval. Four sequences from the J base pool also had half lives recordable within a twenty-four hour time interval (Figure 15). Each sequence exhibited first order rate behavior and single turnover K_{obs} values were collected for all of the individual sequences. The data was collected either by scintillation count or by phosphorimaging (Figure16). The data was then fit to the equation $y=x(1-\exp[-kt])$. The "y" variable is the fraction reacted at time "t." The "x" variable is the fraction reacted at time infinity. The "k" variable is the observed rate constant. This equation can also be transformed so that the data can fit a linear plot and where the kobs value can be attained from the slope, $\ln(x/x-y)=kobs*t$.

For each sequence, a hyperbolic curve was fit to the kinetic data by using least squares approximation. The catalysts were immobilized on a streptavidin support and incubated (23 °C, 50 mM HEPES, 1 M NaCl, 1 mM $MgCl_2$, 0.02 mM EDTA). Nineteen time points were collected from 0-24 hours. The k_{obs} experiments were repeated and were found to have up to 50% variation. However all observed rate constants stayed within the same magnitude.

Figure 15 Unrooted neighbor joining trees using CLUSTAL W and Tree View after 7 rounds of selection The length of the branches are proportional to their estimated divergence. (A) functionalized pool (B) nonfunctionalized pool


A.

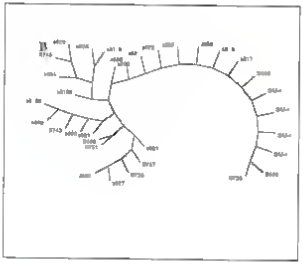


Table 2 The first four columns represent base distributions from the functionalized base selections and the last four columns represents the nonfunctionalized selection.

	A	C	G	T	A	C	G	T
Average	0.22702	0.15337	0.34527	0.23106	0.28515	0.20279	0.25252	0.24864
Standard Deviation	0.03247	0.06926	0.05585	0.04579	0.05276	0.04503	0.03276	0.04554

Figure 16 Functionalized sequences (j 9 j 17b j729 j741), nonfunctionalized sequences (a62 a614b a615b a616) (rA) stands for ribonucleotide.

j79
CTGCAGAATTCTAATACGACTCACTAT(**rA**)GGAAGACATGGCGACTCTCAGGGGJACAJJ
GJGGJJAAAGGJCJGGJACGCCAGAJGJGGJGACGGJAAGCJJGGCAC

j717b
CTGCAGAATTCTAATACGACTCACTAT(**rA**)GGAAGACATGGCGACTCTCGJJGGGCCGCC
AGJCCGCJJAAAGGCAJGGJACGJJCGAJGJGACGGJAAGCJJGGCAC

j729
CTGCAGAATTCTAATACGACTCACTAT(**rA**)GGAAGACATGGCGACTCTCAGGGGJACAJJ
GCGGJCAAAGGJCJGGJACGCCAGAJGJGGTGACGGJAAGCJJGGCAC

j741
CTGCAGAATTCTAATACGACTCACTAT(**rA**)GGAAGACATGGCGACTCTCGGCAJGGGGGG
JAAAAGGCAJACGGGJJACJGAAGJJAJAGJGACGGJAAGCJJGGCAC

s62
CTGCAGAATTCTAATACGACTCACTAT(**rA**)GGAAGACATGGCGACTCTCACACGCACGGA
CTTGTACGTATATAGCGTAAGGTTGATAGTGACGGTAAGCTTGGCAC

s614b
CTGCAGAATTCTAATACGACTCACTAT(**rA**)GGAAGACATGGCGACTCTCACATCATGCGA
GCACACGCAATAGCCTGATAAGGTTGGTAGTGACGGTAAGCTTGGCAC

s615b
CTGCAGAATTCTAATACGACTCACTAT(**rA**)GGAAGACATGGCGACTCTCACACGCACGGA
CTCGCACGTATATAGCGTAAGGTTGATAGTGACGGTAAGCTTGGCAC

s616
CTGCAGAATTCTAATACGACTCACTAT(**rA**)GGAAGACATGGCGACTCTCACTGCACAATC
CAACACCGATTGCTGCAAAGGTTGTTAGGGTGACGGTAAGCTTGGCAC

Figure 17 A DNAzyme exhibiting cleavage of the ribonucleotidyl bond after selection The progress of cleavage was measured in the intervals (minutes) 20, 60, 90, 300, 840 960, 1200, 1380, and 1500 respectively

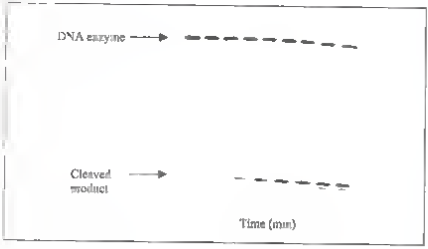

DNA enzyme
Cleaved product
Time (min)

Table 3 Apparent rate constants (k_{obs}) for catalysis from the functionalized and non functionalized libraries

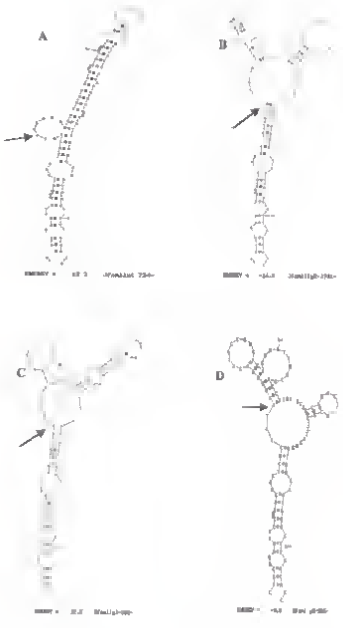
	K_{obs} (min-1)
Standard Pool	
Sequence	
SI2	0 008
SI14	0.003
SI15	0.0035
SI16	0 003
J-Pool	
Sequence	
J79	0 0638
J717	0.0032
J726	0 1
J741	0 0254

Single turnover conditions can done intramolecularly as opposed to intermolecularly because the rate of enzyme substrate association, or the association of two complementary oligonucleotides is several orders of magnitude faster than the rate of catalysis (Santoro, 1998). Thus the rate of cleavage should be the same for an intramolecular cleavage as it is for an intermolecular one. These methods for obtaining k_{obs} have been previously reported (Geyer and Sen, 1997). To determine whether or not these deoxyribozymes could exhibit enzymatic behavior the pool of deoxyribozymes were reacted to an excess of radiolabeled substrate. Over time cleaved product accumulated as expected (Figure 10).

Secondary structural analysis

The secondary structural analysis was performed by submitting known sequences of catalytically active DNA molecules to an online secondary structure prediction program maintained by Dr. Michael Zuker at the University of Washington, (http://mfold.wustl.edu). The software implements energy rules developed by SantaLucia (SantaLucia, 1998). These energy rules determine the energetically optimal folded structure of DNA from a library of DNA duplexes with different conformations. The SantaLucia method uses nearest neighbor thermodynamics and presumes that the stability of a base pair is reliant on the identity as well as the orientation of its neighboring bases. A few secondary structures are represented in Figure 17 from both the functionalized (endowed) and nonfunctionalized (nonendowed) selections.

Figure 18. A&B represent motifs common in the functionalized selectons and C&D represents motifs common in nonfunctionalized selectons The arrow indicates the site of cleavage


A
B
C
D

The secondary structures, for both the functionalized and nonfunctionalized catalysts, exhibit a bulge where the site of cleavage should take place. In the functionalized sequences, it is worth noting that there is a high guanosine nucleobase composition in the bulge. Many of these deoxyribozymes also possess a "pistol like" secondary structure. These motifs have been previously reported in deoxyribozymes capable of hydrolyzing the phosphodiester backbone of RNA (Breaker 1997).

The Quantitative Evaluation of the *In Vitro* Selection: New Mathematical Models for Selection

Several mathematical models are proposed here because there are at present no mathematical formalisms to quantify DNA pool potential for catalysis or mathematical models that describe natural selection and evolution derived from *in vitro* selection experiments. Although *in vitro* selection has created alternative roles for genetic information, little work has been done to answer the general questions about the distribution of these selectable properties (either binding or catalysis) within a set of random nucleic acid sequences(Szostak, 1998). This distribution is crucial for our understanding of the value of *in vitro* selection as a technology for generating new receptors and catalysis. It is also important for assessing the role of *in vitro* selection like processes on the origins of life on Earth or in the universe. By adding protein-like functionality to DNA and quantifying its intrinsic improvement in performing catalysis over nonfunctionalized DNA, we have for the first time been able to answer these theoretical questions with empirical data.

Fitting the Data to both Theoretical and Statistical Models

To begin the quantitative analysis, a probability function is derived that relates the frequency of a selectable property (for example, the rate constant for a reaction to be

catalyzed) in a population to the magnitude of the rate constant. This function should be decreasing, i.e., it is expected that the probability of encountering a good catalyst in a population of random DNA sequences is lower than the probability of encountering a poor catalyst. Further, the definite integral from zero (molecules having no catalytic power at all) to infinity (molecules having outstanding catalytic power) must be bounded (and, traditionally, normalized to unity).

A variety of functions have these properties. For example, an exponential probability distribution may be chosen where the probability of finding a catalyst with a rate constant k, $P(k) = a\exp(-bk)$ with b being a constant that describes the intrinsic limitation of a biopolymer to generate selectable behavior. Different biopolymers, as will be seen below, have different intrinsic ability to generate any particular selectable behavior. The higher the value for b, the faster the function falls towards zero, and this means that there will be more poor catalysts (and the fewer good catalysts) present in a random population of that biopolymer (Figure 18a).

Determining the probability distribution is in principle direct. One needs simply to sample members of a library of random sequences, measure the rate constants for each member sampled, and plot the distribution. In practice, however, this procedure will not be successful with DNA or RNA libraries. In these biopolymers, catalysis is rare, and inefficient when it does occur. Any reasonable sample size identifies no molecules with detectable catalytic behavior. We can, of course, select from the population those molecules that have catalytic activity, and amplify them by the polymerase chain reaction. In this research, the selection system is arranged so that by catalyzing the cleavage, the catalytically active molecule is released from a solid support. If the selection is run for time

t then half of the molecules catalyzing the reaction with a half life of t will be recovered ($k = \ln 2/t$). A larger fraction of the catalysts with a larger k and a smaller fraction of catalysts with a smaller k will also be recovered in the selection. Inactive molecules remain bound to the support.

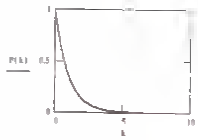
An analytical transformation can be used to describe the selection step. Specifically the selection step multiplies the initial $P(k)$ probability distribution function by a second function, the selection function, $Q(k)$. For the Breaker-Joyce experiment, it is readily shown that the function is analytical: $Q(k) = (1 - \exp(-kt))$ (Figure 18b). This yields a new probability distribution: $P'(k) = P(k)*Q(k)$ (figure 18c). After the selection, amplification using the polymerase chain reaction increases the total number of DNA molecules, but is presumed not to alter the distribution. Subsequent selection steps correspond to a multiplication of the new probability distribution function again by the transformation function: $(1 - \exp(-kt))$, to give yet a new probability. In this example, after n rounds of selection, the probability distribution should have the form: $\exp(-kt)(1-\exp(-kt))^n$. These mathematical transformations can be applied in the inverse. After a sufficient number of selection steps, poor catalysts will eventually be depleted enough to allow detection of catalytically active oligonucleotides. At this point, sampling the population encounters individuals with measurable catalytic power a reasonable fraction of the time, and a probability distribution can be constructed by experiment. A function representing this distribution can then be divided by the selection function n times to generate an approximation of the original probability distribution.

The analysis described above can now be used to describe the data from the selection. The probability distributions for catalysts in the two selection product pools after

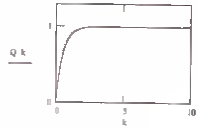
round eight was fitted by a least squares procedure to the "selection equation," $\propto \exp(-bk)(1 - \exp(-kr))$", Figure 19 Although all data from the different rounds of selection fit this distribution round eight was divisible by the most transformation functions (six times for the J/pool and four times for the non-functionalized pool). These yielded an intrinsic potential value of 4 9 and 8 9 for the J containing and standard pools respectively To generate the probability distribution P (k) this function was then divided by the transformation function n times These functions are shown in Figure 20b

Figure 19 Probability of finding catalysts as a function of rate. C is inversely proportional to the intrinsic potential for catalysis. k is the rate constant for the catalytic reaction and t is the time allowed by the selection procedure for catalysis to occur. (A) the distribution of catalysts before selection (B) the selection function (C) the new distribution of catalysts after the initial distribution of catalysts are transformed by the selection function

A. $P(k) = Ce^{-Ck}$



B $Q(k) = 1 - e^{-kt}$



C. $F(k) = Ce^{-Ck}\left(1 - e^{-kt}\right)$

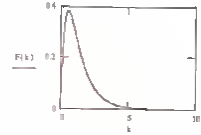


Figure 20 Round 8 functionalized (red) and nonfunctionalized (blue) data fit to the equation $a\exp(-bk)(1-\exp(-kt))$

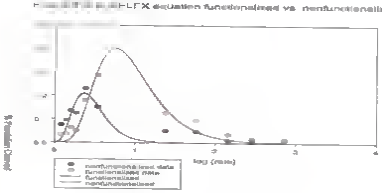
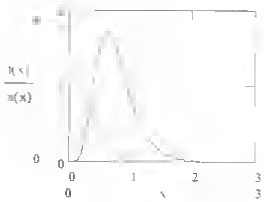
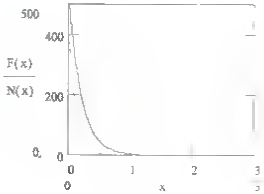

log (mass)
nonfunctional test data
functionalized data
functionalized
nonfunctionalized

Figure 21 (A) Round eight selector functionalized (red) and nonfunctionalized (blue) fit to the equation $C\exp(-Ck)(1-\exp(-kt))^n$ (B) Round eight selection functionalized (red) and nonfunctionalized (blue) divided n times by the transformation function $1-\exp(-kt)$

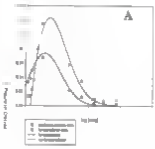

N(x)
n(x)
0
0
0
1
2
3
0
x
3

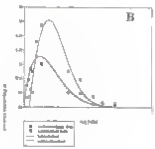

500
400
F(x)
N(x)
200
0
0
0
1
2
3
0
x
3

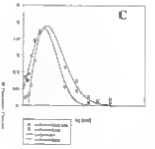
At the outset, an exponential probability distribution was assumed for P(k) as well as F(k) (Figures 19a and 19c). This assumption was arbitrary. Should we wish to relax this assumption, we can fit the empirical data obtained after n rounds of selection to any numerical function, and transform this using the selection function to yield a numerical representation of the initial probability distribution that need not be exponential. To illustrate this, the probability distribution data for catalysts in the two selection product pools after rounds seven through eleven was fitted by a least squares procedure to a Weibull distribution (Weibull, 1951), Figure 21.

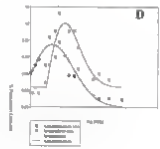
The Weibull distribution is frequently used in statistics as a generic function for representing data requiring the simultaneous performance of several components. In this example, catalysts require the simultaneous action of several different functional groups to act as general acid or base catalysts as well as multiple binding interactions. It is apparent from the Weibull fit to the selection rounds, that the functionalized pools possess faster catalysts at each round of selection. At rounds ten and eleven, the selection pressure was tightened to five minutes as opposed to two hours. A noticeable shift to the right in both functionalized and nonfunctionalized pools is observed (table 4). As done with the "selection function," the Weibull distribution for round eight was divided n times by the transformation function to yield the probability distribution of the initial population (Figure 22d). The overall appearance of P(k)Weibull is not markedly different from P(k)exponential. It is different theoretically of course, at the limits of extremely fast and extremely slow catalysts, but these differences do not fall outside of the error in the present data.

Figure 22 Rounds 7-11 fit to a four-parameter Weibull distribution. The red line represents the functionalized fit and blue represents nonfunctionalized. (A) round 7, (B) round 8, (C) round 9, (D) round 10, and (E) round 11


A


B


C


D

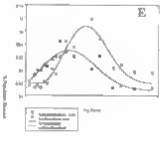

E

Figure 23 (A) The Weibull distribution. *c* is the shape parameter. *b* determines the range, *a* determines the capacity, and *x0* determines where the function begins on the x-axis. (B) The Weibull distribution divided *n* times by the transformation function 1-exp(kt). (C) Round 8 data fit to the Weibull distribution for functionalized (red) and nonfunctionalized (blue) selections. (D) The Weibull distribution for functionalized (red) and nonfunctionalized (blue) divided eight times by the transformation equation 1-exp(kt) to show the original distribution

A

B

C

D

Quantifying Improvement Using the Selection Equation

From a qualitative (not quantitative) perspective a theoretical fit from the selection equation and a statistical fit by the Weibull distribution have both shown that the initial distribution profiles for functionalized DNA possess better catalysts. Several other qualitative indications also lean towards evidence that a functionalized DNA pool is intrinsically better. For example, although the none of the data from the selection rounds were able to fit the selection equation the functionalized pool consistently fit the selection equation with a higher number of transformation functions ($1-\exp^{-x}$). This may imply that theoretically, the functionalized pool was enriched with catalysts in earlier rounds of selection. There could have been more faster catalysts in the functionalized pool. However this was not obvious from our empirical data. The detectable elution of catalysts from both the functionalized DNA and nonfunctionalized DNA occurred at the same time. But a higher percentage of catalysts were consistently eluted from the functionalized pool (Figure 12) and greater number of catalysts could be the reason.

The average rates among the separate pools after each round of selection also yielded some qualitative data (table 4). Not only was the average rate better for the functionalized pool in each round of selection, but the slowest observed rate for the functionalized pool (0.0067) was still faster than the fastest observed rate for the nonfunctionalized pool (0.0027) pool. This is no expected since the selection pressure for rounds ten and eleven were tightened to five minutes as opposed to two hours.

Because we have defined the intrinsic potential to mean an improvement in the probability of finding catalysts, a direct correlation can be made between the average rate found at a selection round and the intrinsic potentials derived from that selection round.

Theoretically we can make this quantitative statement here, but unfortunately the failure of the data to be fit by an equal magnitude to its selection round makes this quantitative statement analytically improbable.

There is also the matter of the intrinsic potential. The intrinsic potential is affected by the variable in front of the exponent otherwise known as the *carrying capacity*. If the probability distribution for an exponential function is set to unity then by definition both constants must be the same (Nelson 1982). On the other hand, it is unreasonable to assume that the behavior of this system can be defined by only one variable. That is chemical reactions involve a series of different interactions that can be described by many different variables. Still, the modified "selection equation" equation does fit all the data for all the selection rounds and this statement though weak, still holds some qualitative (but not quantitative) validity.

Table 4 The average rate of catalysis of functionalized and nonfunctionalized pools after each round of selection As with the individual sequences the kobs was determined by nonlinear regression analysis and scintillation count

	Round 7	Round 8	Round 9	Round 10	Round 12
Functionalized kobs (min^-1)	0.0067	0.0071	0.0068	0.0078	0.0103
Nonfunctionalized kobs (min^-1)	0.0007	0.0008	0.0013	0.0022	0.0027

3. Quantifying Improvement Using the Weibull Distribution

A better way of quantifying improvement in the intrinsic potential for a functionalized library to generate catalysts makes reference to the Weibull distributions. The Weibull distribution is ideal because it uses several parameters. These appear to be needed to capture the features of the probability-catalyst distribution. The data from each round of selection fit a Weibull distribution with high R squared values.

Dividing the Weibull distributions describing the data from the functionalized library by the Weibull describing data from the nonfunctionalized library in round 7 generates a quotient curve, shown in Figure 24. The nonfunctionalized pool has more poor catalysts than the functionalized pool. Because we are dividing the functionalized Weibull (numerator) by the nonfunctionalized Weibull (denominator), the ratio should be less than one for low values of k_{obs}. For faster catalysts, the denominator becomes smaller, the numerator becomes larger, and the curve approaches unity. For still faster catalysts, the functionalized Weibull is larger than the unfunctionalized Weibull, and the ratio becomes greater than unity and continues to increase.

The quantitative improvement of functionalization can be measured by looking at quotient curve or divisitory product (curve) from round seven. Round seven is ideal because it is the first round of selection in which catalytic DNA appears. We can arbitrarily take the mode (the peak of the quotient curve) for Round 7 or any value (rate) in the portion of the distribution greater than one. We can then determine a probability of finding this rate, say X, for both the functionalized and nonfunctionalized distributions. For the nonfunctionalized distribution we can estimate a probability Y of finding this rate. We can do the same for the functionalized distribution and estimate a probability Z.

With these probabilities, a quantitative statement can be made where a rate of X is Z/Y times more likely to be found in a functionalized pool than a nonfunctionalized pool.

The probability of find a catalyst with a k_{obs} greater than E is defined as $P(E) = \int_E^{\infty} f(y)dy$ where f(y) is the distribution function. To determine the probability of finding all the catalysts up to rate X we use the cumulative distribution function $F(y) = P\{Y \leq y\}$. On the other hand, to determine the probability of finding all the catalysts that are rate X and faster, the reliability function, $R(y) = P\{Y>y\} = 1 - F(y)$, would be used. For the Weibull distribution, F(y) is defined as $1-\exp[-(y/\alpha)^{\beta}]$, where α is the scale parameter and β is the shape parameter. This applies to any parameter Weibull distribution (Nelson 1982, Johnson 1994).

Before drawing conclusions about the improvement in the catalytic potential of a DNA library arising through functionalization, the raw data must be put through a rigorous statistical test to assess the possibility of differences due to stochastic uncertainty. One well-known method of statistical testing involves calculating the average rates and their respective 95% confidence intervals. If the confidence intervals do not overlap, then the averages are statistically significant. At round seven an average rate of 0.0014/min with a 95% confidence interval of 0.0011 was calculated for the functionalized pool, and an average rate of 0.0004/min with a 95% confidence interval of 0.00029/min was calculated for the nonfunctionalized pool. Clearly the average rates and their confidence intervals do not overlap. This indicates that the difference between the functionalized pool and nonfunctionalized pool is statistically significant. We can now say with some confidence that a quantitative estimation of improving the probability

in finding catalysts (or any subsequent transformation of this data) will be statistically significant.

At round seven, the mode of the moment curve was found to be 0.0047/min. Using the reliability function R(y), the probability of finding a catalyst with a rate of 0.0047 or faster for the functionalized pool at round seven was 0.57 (Figure 25a). The probability of finding the same catalyst or faster in the nonfunctionalized pool was only 0.42 (Figure 25b). The following quantitative statement can now be made. The functionalized pool is 1.36 times more likely to find a catalyst with a rate of 0.0047/min or faster than the nonfunctionalized pool.

Figure 24. The quotient curve for round 7 functionalized and nonfunctionalized data. The data for the functionalized selection was divided by the data for the nonfunctionalized selection. The result of the quotient was then fit to a Weibull distribution.

Round 7 Quotient Curve Dividing Endowed by Nonendowed Weibull Distributions

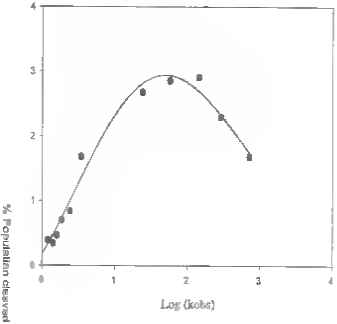


Figure 25 The reliability function R(y) calculated for the functionalized pool and nonfunctionalized pool at round seven selection For the functionalized pool b7=0.9152 c7=1.8707 x70=0.7717 and x=0.828 For the nonfunctionalized pool b71=0.1838 c71=1.639 x701=0.6093 and x701=0.6093

A

$$e^{-\left[\left(\frac{x-x70}{b7}\right)+\left(\frac{c7-1}{c7}\right)^{\frac{1}{c7}}\right]^{c7}}=0.576$$

B

$$e^{-\left[\left(\frac{x-x701}{b71}\right)+\left(\frac{c71-1}{c71}\right)^{\frac{1}{c71}}\right]^{c71}}=0.423$$

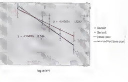
Evaluating the Affect of the J-base

As an alternative way to determine how the presence of the J-base contributed to the improvement of, the probability distribution data were plotted using a log-log (Figure 20). The ratio of the slopes between endowed pool over the unmodified DNA pool was approximately 1.5. The slope of the line describing catalysis by the functionalized pool was 0.4, the corresponding slope was 0.6 for the nonfunctionalized pool. The slopes may be used to estimate the number of positions in the nucleotide sequence that must be defined to obtain a factor of 10 increase in the rate of catalysis. Specifically, to improve the rate of catalysis for the nonmodified pool by a factor of 1000 three additional positions in the nucleotide sequence must be defined. In contrast, two nucleotides in the sequence must be defined to improve the rate of catalysis in the functionalized DNA pool by a factor of 1000.

Analysis of variance was determined from the two sets of data using the software Statistical Analysis Software (SAS). This suggested that the differences in the curves is statistically significant. A p-value of 0.0025 was derived, thus rejecting the null hypothesis that both sets of data fit better to one curve than to two separate ones.

Figure 26 Functionalized versus nonfunctionalized linear regression curve after seven rounds of selection using log transformation.

Functionalized vs. Standard Base Pool Catalysts

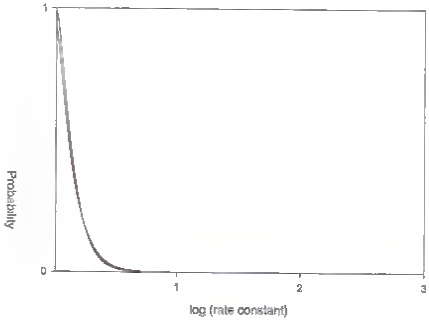


The previous Weibull distribution analysis estimated the improvement of the functionalized library over the nonfunctionalized library by comparing data at round seven without extrapolation back to round zero. This estimation is adequate because any transformation of these data would still reflect the properties of the distribution for that particular round, whether it is mutation, selection pressure, or experimental conditions. Still, the extrapolated data to round zero would represent the initial distribution to a close approximation. We can also visually approximate the initial distribution of catalysts for both libraries as a function that is exponential-like. As we will observe, the quantitative improvement by extrapolating back to round zero and comparing directly from round seven will give the same value of improvement for the functionalized library.

To obtain these initial distributions, the Weibull functions of round seven for both the functionalized and nonfunctionalized libraries were divided seven times by the transformation function, $Q(k) = 1 - \exp(-kt)$ (figure 28). The distributions revealed that improvement for the functionalized data is (again) approximately 1.4 times over that for the nonfunctionalized data. Specifically, the functionalized pool is 1.4 times more likely to find catalysts with rate constants 0.001 per minute than the nonfunctionalized pool. The fastest catalysts, according to these distributions, are 0.06 per minute for the functionalized pool and 0.03 per day for the nonfunctionalized pool.

Figure 28 Weibull functions of round seven for both the functionalized and nonfunctionalized libraries divided seven times by the transformation function, $Q(k) = 1 - \exp(-kt)$

Round Zero Probability Distribution of Catalysts Functionalized vs Nonfunctionalized



Pages
Missing
or
Unavailable

MATERIALS AND METHODS

Oligonucleotides and Nucleotide Analogs

The starting pool and primers used for all experiments were synthesized by Integrated DNA Technologies Inc. and PAGE purified. The starting pool, 5 -GTGCCAAGCTTACCGTCG-N40-AGATGTCGCCATCTCTTCCTATAGTGAGTCGTATTAG-3 was ordered to have the N40 random region to have an average of nucleotide distribution of a ratio 1:1:1:1. Primer 1 5 CTGCAGAATTCTAATACGACTCACTATAGGAAGACATGGCGACTCTC-3 came in two forms. Primer 1 was either biotinylated on the 5' end for binding to the streptavidin column for selection or it was 5 gamma-$AT^{32}P$ labeled for PAGE analysis. Primer 3 5'-GTGCCAAGCTTACCGTCAC 3 was either not modified or it was in the 5 biotinylated form to make single stranded DNA.

The J base triphosphate was synthesized by Simona Jurczyk and Tom Battersby biosciences of EraGen of Alachua, Florida and purified by HPLC. The protected J base was also synthesized by EraGen and purified by HPLC. Both compounds were analyzed by proton and phosphorous nuclear magnetic resonance.

DNA Pool Preparation

Round zero for both the standard base selection and the J-base selection were prepared by PCR using the 80 pmoles of starting pool. The PCR reaction was done with 0.5 units/μL of Vent polymerase (New England Biolabs Inc.) in the presence of 10 mM KCl, 10 mM $(NH_4)_2SO_4$, 20 mM Tris-HCl (pH 8.8), 2 mM $MgSO_4$, 0.1% Triton X-100, 60 pmoles of primer 1 (biotinylated), and 60 pmoles of primer 2 with no modification, and 0.25 mmoles of each dNTP. Thymidine triphosphate was used only for the standard base selection and JTP was only used for the J-base selection, at no time was either nucleotide triphosphate interchanged during selection. The PCR was cycled at 94°C for 2 minutes, 55°C for 2 minutes, then 72°C for 4 minutes for eight cycles. Each PCR reaction was done in 200 μL. The PCR product was then phenol chloroform extracted twice and then ethanol precipitated with glycogen as a carrier. The ethanol precipitation was carried overnight at -80°C and subsequently centrifuged at 4°C at 14,000 RPMs for 20 minutes. The supernatant was decanted and 70% ethanol was added to each sample. The sample was then centrifuged for 2 minutes at 14,000 RPMs at 4°C. The supernatant was decanted and the samples were placed in the Eppendorf speed vacuum for 45 minutes. When dry, the samples were stored at 4°C until selection was ready to be performed.

***In vitro* selection**

Each of the two starting pools was redissolved in 200 μL of Tris-EDTA (10mM Tris HCl, 1 mM EDTA) (pH 7.6). One hundred microliters was stored at -20°C and the rest were used for selection. Two 15 mL Bio Rad chromatography

polystyrene columns were labeled and 200 µL of streptavidin (Fluka Biochemicals) were loaded to each column. Each column was washed with 1 mL of Tris-EDTA buffer (10mM Tris HCl 1 mM EDTA) (pH 7.6) and capped. The redissolved pool was added to the column and gently agitated to resuspend the streptavidin with the pool mixture. The column and pool were incubated at 23°C for 20 minutes. The column was then washed 10 times with TE (pH 7.6) followed by 3 volumes of 0.2 M NaOH, then neutralized with 8 column volumes of Buffer A (50 mM HEPES 1 M NaCl pH 7.6). One column volume of reaction buffer or Buffer B (50 mM HEPES 1M NaCl, 1 mM $MgCl_2$, 0.02 mM EDTA) was added to equilibrate the immobilized DNA with the divalent cation. The single stranded DNA was eluted for 20 minutes, then 40 minutes then 2 hours with 200 µL of Buffer B after each successive incubation. The eluted filtrate was ethanol precipitated, 3 M ammonium acetate, and glycogen as a carrier at -80°C overnight. The ethanol precipitation was centrifuged at 4°C at 14,000 revolutions per minute (RPMs) or 12,500g for 20 minutes. The supernatant was decanted and 70% ethanol was added to each sample. The sample was then centrifuged for 2 minutes at 14,000 RPMs at 4°C. The supernatant was decanted and the samples were placed in the Eppendorf speed vacuum for 45 minutes.

The samples were resuspended in 100 µL of double distilled water (Millipore Systems). Test PCR reactions using 5 µL of resuspended sample added to 50 µL of PCR mix (60 pmoles primer 1 and primer 2, 0.25 mmoles dNTP, 2 mmoles $MgSO_4$, 10X buffer) and 0.5 units/µL of Vent DNA polymerase were done until a single band could be visualized on 4% agarose gel. Usually 15-20 cycles were done. Using the same conditions large scale PCRs were done. Four of the PCR reactions for the

large scale amplifications were radiolabeled with alpha $dAT^{32}P$. The PCR product was then phenol-chloroform extracted twice and then ethanol precipitated with glycogen as a carrier. The ethanol precipitation was centrifuged at 4°C at 14,000 RPMs (12,500g) for 20 minutes. The supernatant was decanted and 70% ethanol was added to each sample. The sample was then centrifuged for 2 minutes at 14,000 RPMs (12,500g) at 4°C. The supernatant was decanted and the samples were placed in the Eppendorf speed vacuum for 45 minutes.

Rounds 1-4 were performed as above. rounds 5-12 were done using 0.5 units/μL of Vent exo- polymerase to promote evolution via error prone PCR. Rounds 10-12 were collected after 5 minutes.

Cloning and Sequencing

For subcloning and sequence analysis, molecules eluted at the end of rounds five, six, and seven selection were amplified by PCR using Taq DNA polymerase to reincorporated standard TTF deoxynucleotides instead of JTF deoxynucleotides in the functionalized pool. PCR was performed under the following conditions: 100 μL PCR reactions (60 pmoles primer 1, 60 pmoles primer2, 0.25 mmoles dNTPs, 50 mM Tris-HCl, 100 mM NaCl, 0.1 mM EDTA, 1 mM DTT, 50% glycerol and 1% Triton, 1.5 mM $MgCl_2$). 8-15 cycles of 1 minute at 94°C, 2 minutes at 55°C, and 2 minutes at 72°C with a final extension for 8 minutes at 72°C. The final extension allowed Taq to add a single contaminated deoxyadenosine for cloning into a vector with a deoxythymidine 3' overhang. The pCR 2.1-TOPO vector (Invitrogen), a pUC 18 derivative, was used for cloning. After a five minute ligation, competent E. coli strain

TOP 10 (Invitrogen) was transformed as described by the supplier and plated onto LB plates with kanamycin. IPTG was used for blue-white screening.

White colonies were randomly chosen and streaked onto LB plates with kanamycin. Plasmid purification was performed by traditional alkaline lysis methods or by using a Qiagen kit. Sequencing reactions used the Big Dye Terminator (Applied Biosystems, Inc.). PCR conditions for the sequencing reaction were as follows: 25 cycles of 30 s at 96°C, 15 s at 50°C, and 4 minutes at 60°C. PCR products were held at 4°C and then applied to handmade Sephadex G-50 spin columns made from 0.5 mL eppendorf tubes and 1/8 inch polyethylene filter material (Porex). Purified plasmids were submitted as templates to a sequencing service (Sequencing Core, University of Florida).

Plasmid Miniprep

Approximately 3-5 mL overnight cultures were grown from individual colonies in Terrific Broth (Sigma) or LB Broth (Difco) in 15 mL Falcon tubes in a 37°C shaker. Three mL of culture were pelleted in microfuge tubes by centrifugation at 14000 RPM for 3 minutes. The media was then decanted. The bacterial pellet was resuspended in 300 µL glucose/Tris/EDTA (GTE) solution with RNase A by gentle vortexing, making sure that the entire pellet was resuspended. The cells were lysed using 300 µL SDS/0.2 N NaOH and inverted several times until the solution cleared. Three hundred microliters of neutralization solution (5 M potassium acetate) was added, with gentle inversion. The resulting precipitate was centrifuged for 10

minutes at 14 000 RPMs to pellet denatured proteins and chromosomal DNA. The supernatant was then transferred to sterile microfuge tubes.

Chloroform extraction was done to remove any remaining proteins. The upper phase was transferred to clean microfuge tubes and precipitated with propanol and centrifinging for 30 minutes. The supernatant was carefully decanted and discarded. The pellet was washed in 70% ethanol and vortexed. After briefly centrifinging to reposition the pellet, the supernatant was decanted, and the remaining liquid was carefully aspirated. The pellets were air dried or heated briefly (5 seconds) on a 75°C heat block. Dried pellets were resuspended in 50 µL ddH_20. Fifty microliters of PEG8000/1.6 M NaCl was added and allowed to incubate on ice for 30-60 minutes to remove salts. After a 15 minute centrifugation at 12 000 RPMs to pellet the plasmid DNA, the pellet was washed with 500 µL 70% ethanol, vortexed, recentrifuged briefly, decanted and carefully aspirated. The pellets were dried again on a 75°C heat block.

Primer Extension Screening

Both primer 3 (5'GGTCGTCTAGAGTATGCGGTAG-3' and Template 1 (5'ACCCTACCGCATACTCTAGACGACC-3') were purchased from Integrated DNA Technologies and PAGE purified. Primer 3 was redissolved in double distilled water to have a final concentration of 1microgram/microliter. In a final reaction volume of 50 µL, 33 µL of primer was added to 5 µL of 10X buffer (70 mM Tris HCl pH 7.6, 10 mM $MgCl_2$, 5 mM dithiothreitol), 2 µL of gamma $AT^{32}P$ (Pharmacia Biotech), 2 µL of T4 polynucleotide kinase (New England Biolabs) and 8 µL of

double deionized water. The kinase reaction was incubated at 37°C for 30 minutes and subsequently phenol-chloroform extracted and then ethanol precipitated. Both the primer and template were redissolved in double deionized water to have a final concentration of 3 pmol/μL. Overall, 10 polymerase were screened (5 family A and 5 family B polymerases). Each reaction volume had a total volume of 30 μL. In each reaction there were 5 μL of primer and 7 μL of template. The primer-template solution was heated for 3 minutes at 94°C and allowed to cool to room temperature. The reactions were set up so that each polymerase screened used either the four standard nucleotides or three standard nucleotides plus dTP instead of TTP.

For the family B polymerases the reaction conditions were: 1. Vent exo- 12 μL primer template solution, 0.7μL of standard 10mM dNTPs or 0.7μL 10mM dTP, 3 μL 10X buffer (10mM KCl, 10mM $(NH_4)_2SO_4$, 20 mM Tris HCl (pH 8.8), 2 mM $MgSO_4$, 0.1% Triton X-100), 1 unit of Vent exo- polymerase, and 11.7μL of ddH_2O. The reaction was incubated at 72°C for four minutes. 2. Deep Vent, 12 μL primer-template solution, 0.7 μL of standard 10 mM dNTPs or 0.7 μL 10 mM dTP, 3 μL 10X buffer (10 mM KCl, 10 mM $(NH_4)_2SO_4$, 20 mM Tris-HCl (pH 8.8), 2 mM $MgSO_4$, 0.1% Triton X-100), 1 unit of Deep Vent polymerase, and 11.7 μL of ddH2O. The reaction was incubated at 72°C for four minutes. 3. T4, 12 μL primer template solution, 0.7μL of standard 10 mM dNTPs or 0.7μL 10 mM dTP, 3 μL 10X buffer (50 mM NaCl, 10 mM Tris-HCl, 10 mM $MgCl_2$, 1 mM DTT, pH 7.5), 1 unit of T4 DNA polymerase, 1.5 μL of bovine serum albumin (BSA), and 10.2 μL of ddH_2O. The reaction was incubated at 37°C for four minutes. 4. Exo- Pfu, 12 μL primer-template solution, 0.7μL of standard 10 mM dNTPs or 0.7 μL 10mM dTP, 3 μL 10X

buffer (200mM Tris-HCl pH 8.8, 100 mM KCl, 100 mM $(NH_4)_2SO_4$, 20 mM $MgSO_4$, 1% Triton X-100, 1 mg/mL nuclease-free bovine serum albumin), 1 unit of Exo- Pfu, and 11.7 µL of ddH_2O. The reaction was incubated at 72°C for four minutes. 5. Pwo: 12 µL primer-template solution, 0.7 µL of standard 10 mM dNTPs or 0.7 µL 10 mM JTP, 3 µL 10X buffer (100mM Tris-HCl pH 8.85, 250mM KCl, 50 mM $(NH_4)_2SO_4$, 20 mM $MgSO_4$), 1 unit Pwo, and 11.7 µL ddH_2O. The reaction was incubated at 72°C for four minutes.

For the family A polymerases the reaction conditions were: 1. Bst: 12 µL primer-template solution, 0.7 µL of standard 10 mM dNTPs or 0.7 µL 10 mM JTP, 3 µL 10X buffer (20 mM Tris-HCl, 10 mM KCl, 10 mM $(NH_4)_2SO_4$, 2 mM $MgSO_4$, 0.1% Triton X-100, pH 8.8), 1 unit Bst, and 11.7 µL ddH_2O). The reaction was incubated at 72°C for four minutes. 2. T7: 12 µL primer-template solution, 0.7 µL of standard 10 mM dNTPs or 0.7 µL 10 mM JTP, 3 µL 10X buffer (20 mM Tris-HCl, 10 mM $MgCl_2$, 1 mM DTT, pH 7.5), 1 unit T7, and 11.7 µL of ddH_2O. The reaction was incubated at 72°C for four minutes. 3. Taq: 12 µL primer template solution, 0.7 µL of standard 10 mM dNTPs or 0.7 µL 10 mM JTP, 3 µL 10X buffer (50 mM KCl, 10 mM Tris HCl, pH 9.0, and 0.1% Triton X-100), 1.8 µL $MgCl_2$, 1 unit Taq, and 9.9 µL of ddH_2O. The reaction was incubated at 72°C for four minutes. 4. Tth: 12 µL primer template solution, 0.7 µL of standard 10 mM dNTPs or 0.7 µL 10 mM JTP, 3 µL 10X buffer (50 mM KCl, 10 mM Tris-HCl, pH 9.0, and 0.1% Triton), 1 unit Tth, 3 µL $MgCl_2$, and 8.7 µL of ddH_2O. The reaction was incubated at 72°C for four minutes. 5. DNA polymerase I: 12 µL primer template solution, 0.7 µL of standard 10 mM dNTPs or 0.7 µL 10 mM JTP, 3 µL 10X buffer (10 mM Tris-HCl, 5 mM

$MgCl_2$, 7.5 mM DTT pH 7.5), 1 unit DNA pol I, and 11.7 µL of ddH_2O. The reaction was incubated at 72°C for four minutes.

The reaction mixtures were quenched with equal amounts of formamide/EDTA loading buffer (9 mL formamide, 0.1 mL 0.5M EDTA, 7 µg bromophenol blue (Fisher Scientific)). Seven microliters were loaded on to a 12% PAGE denaturing gel (1 X TBE, 2000 V, 3 hours), dried for 20 minutes at 80°C on a gel dryer (BioRad), and visualized on a phosphorimager (BioRad).

Determining Microscopic Rate Constants for Family B Polymerases

For the running start experiments, the following oligonucleotides were used: Primer (5'GGTCGTCTAGAGTATGCGGTAG) and the Template (5'ACCCTACCGCATACTCTAGACGACC). Both DNA oligonucleotides were PAGE purified and purchased from Integrated DNA Technologies. The following reagents were prepared: Quenching solution (0.02 M EDTA in 95% Formamide), primer/template stock solution (97µL 290 nM 5' radiolabelled Primer: 133 µL 250 nM template, 20 µL 10X buffer: 10 mM KCl, 10 mM $(NH_4)_2SO_4$, 20 mM Tris-HCl (pH 7.5), 2 mM $MgSO_4$, 0.1% Triton X-100); trap stock solution: (75 µL 100 mM unlabelled primer, 150 µL of 100 mM template stock, 25 µL of 10X buffer). Polymerase stock solution (diluted to a concentration 0.1 units/µL), and dGTP or running base stock solution (25 µM). The primer/template and trap stock solutions were heated to 94°C for 3 minutes and allowed to cool to room temperature within 30 minutes. Both solutions were then added together.

Concentrations of 0 μM, 0.1 μM 1 μM 2 μM, 3 μM, 7μM, 14μM, 27.5μM target TTF and JTF solutions were made after adding 10 μL of polymerase stock solution and 12.5 μL of GTP stock solution The overall volume of the target dNTF solution was 50 μL.

The reactions began by placing 5 μL of the primer/template/trap solution into different labeled microfuge tubes according to dNTP concentration. The 5 μL of the appropriate concentration of target dNTF was added The reaction mixture was placed on a heating block set at 72°C and incubated for 5 minutes. The reaction was then quenched with 10 μL quenching solution and placed on ice. The samples (7μL) were loaded onto a 12% PAGE gel and ran at 2500V The gels were dried and exposed to phosphorous isotope imaging plates and visualized and quantified on the Molecular Imager (Bio Rad).

dU base Pool versus Standard Base Pool Intermolecular Cleavage

Two Bio Rad chromatography columns were labeled and 200 μL of streptavidin (Fluka Biochemicals) were loaded to each column. Each column was washed with 1 mL of Tris EDTA (10mM Tris HCl 1 mM EDTA) buffer (pH 7.6) and capped. The redissolved pool was added to the column and gently agitated to resuspend the streptavidin with the pool mixture The column and pool were incubated at 23°C for 20 minutes. The column was then washed 10 times with TE (pH 7.6), followed by 3 volumes of 0.2 M NaOH The eluted NaOH fraction was neutralized with 3 column volumes of 0.2 HCl The single stranded DNA was ethanol precipitated with glycogen as a carrier and incubated at -80°C overnight. The

ethanol precipitation was centrifuged at 4°C at 14,000 RPMs for 20 minutes. The supernatant was decanted and 70% ethanol was added to each sample. The sample was then centrifuged for 2 minutes at 14,000RPMs at 4°C. The supernatant was decanted and the samples were placed in the Eppendorf speed vacuum for 45 minutes.

The pellet was redissolved in 100 µL neutralization buffer (50mM HEPES, 1M NaCl, pH 7.6) and measured by UV absorption at 260 nm. Thirty pmoles of pool DNA from both the standard and nonfunctionalized pools were taken out and added to two separate Eppendorf tubes containing 120 pmoles of substrate (5'CTGCAGAATTCTAATACGACTCACTATrAGGAAGACATGGCGACTCTC-3') that was redissolved in 99 µL neutralization buffer. The reaction mixture was initiated by adding 1 µL of 0.1M $MgCl_2$. The reaction was allowed to proceed for 2 hours at room temperature and quenched with 0.02 M EDTA in 95% formamide. Ten microliters were loaded onto 12% PAGE and run at 2500V for 2.5 hours. The gel was dried and exposed overnight to ^{32}P exposure plates (Bio Rad). The bands were visualized with a phosphoimager (Bio Rad).

Kinetic Analysis of Standard Base Sequences and J-base Sequences

Seven standard and seven J pool sequences were tested for catalytic activity. Sequences were chosen to represent the several catalytic families that had emerged by round seven of selection. The standard sequences and complements of the J-pool sequences were synthesized (Integrated DNA technologies), biotinylated, and radiolabeled. The J-pool sequences were PCR amplified using a 3' biotinylated primer which contained the single ribonucleotide in a 100-µL reaction (5 U µL^{-1}

PWO polymerase, 10X buffer containing 20 mM $MgSO_4$, 0 2 mM dNTP, 007mCi α-dATP). The PCR was set to run for15 cycles at 60 s for 94°C, 120 s for 55°C, and 120 s for 72°C

Following ethanol precipitation the catalytic activity of each DNA was investigated under SELEX conditions. Over a 24-hour period one column volume of reaction buffer was collected at each of the time-points. Fraction of precursor DNA cleaved versus time was plotted from either scintillation count data or quantified by the Molecular Imager (Bio Rad) The catalytic rates were obtained by nonlinear regression analysis (Microsoft Excel and Sigma Plot) using the function $y=x(1-e^{-kt})$ where y represents fraction x represents the cleavage at time infinity and k represents the observed rate constant.

CONCLUSIONS

As stated before, the goal of this research is to create a better candidate for the single biopolymer life form hypothesis. Not only were catalysts created from functionalized DNA, but a chemically endowed DNA library has been shown to possess a greater potential for producing catalysts over a nonendowed DNA library. These results have alluded to two novel concepts in the fields of *in vitro* selection and the origins of life theories.

Use of a Functionalized Base by a Thermophilic Polymerase (First Concept)

The first general concept can be stated: *A protein like DNA triphosphate base can be PCR amplified by a thermophilic Archaeal polymerase. This amplification is efficient enough to be used for in vitro selection experiments to derive multifunctional oligonucleotides for catalysis.*

Specifically, a functionalized thymidine with a positive charged amine group on the 5-position can be incorporated into a template by Vent polymerase with a decreased relative velocity of approximately 30-50% compared to the standard TTP substrate. When the positive charge on the I base is protected with a phenylacetyl group, the relative Vmax is restored, indicating that the positive charge is responsible for the decrease in the rate of polymerization. However, the ability of Vent polymerase to bind to this molecule is only moderately affected, and the Km values between the TTP substrate and the ITP molecule can nearly be corrected within one standard deviation.

An evolutionary statement has also been made. A screen between family B and family A polymerases show that all of family B polymerases can use the J-base as a substrate, but most of family A cannot. A specific aspartate residue, analogous to the Vent D543 in the family B polymerase is suspected to interact with the amine portion of the J-base. The sequence alignment presented in this research give proof that all family B polymerases have this aspartate residue while family A polymerases do not. Further, the computer generated structure of the Vent polymerase with a J-base substrate (modeled by threading a polymerase with over 80% sequence identity and using energy minimization methods) show that the amine propyuyl arm of the J-base is at or in close proximity to the aspartate residue 543. This is the first time a three-dimensional structure is proposed for the Vent polymerase and a modified substrate has been used to probe the active site to determine an evolutionary predilection. The following are postulates derived from the aforementioned findings concerning the first general concept.

Postulate 1. When using a family B polymerase, if the functionalized nucleotide substrate has a Km correctable within one standard deviation compared to its standard nucleotide counterpart, and its Vmax is only moderately affected, then all other Family B polymerases are capable of catalyzing a phosphodiester bond between the functionalized base and its primer.

Postulate 2. Although the positive charge from the J-base decreases the Vmax, it's presence also allows the family B polymerases to incorporate the J-base triphosphate by drawing the positive charged functional group away from the two other aspartate residues in Region I and Region II (in Vent these residues are D407 and D545) that are more important for catalysis. The absence of this analogous residue in family A polymerases puts these family of polymerases at a disadvantage in incorporating the dJ triphosphate.

From the J-base sequence data, it is clear that the J-base is not selected against and is found in the conserved region. The original pool of DNA was randomized such that 25% of each base should be represented in most of the sequences. In both the J-base

and standard selection, the J or T content is at 25% in the random regions. This leads us to the third postulate:

> *Postulate 3*: If the Km of the functionalized nucleotide is correctable within one standard deviation to its standard nucleotide counterpart, and its Vmax is only moderately affected, then the functionalized nucleotide will not (necessarily) be selected against during polymerase chain reaction amplification.

To recapitulate: the first general concept has generated three postulates that validate the methodology for using functionalized bases for the purposes of PCR amplification and in vitro selection. Specifically, we have (a) addressed the relationship between functionalized nucleotides and different families of polymerases, (b) addressed how and what mechanisms these relationships are based, and (c) addressed whether or not these relationships makes it a viable tool for *in vitro* selection. By validating the methodology used to pursue the hypothesis, we can now move on to the hypothesis itself and proceed to the second general concept.

Functionality Improves Catalysis (Second Concept)

The second general concept focuses on the *in vitro* selection and evolution experiments done to create the catalytically active functionalized DNA molecules. The second general concept can be stated: *The addition of functionality (or added chemical groups) to DNA improves the probability of finding better catalysts.*

Using the Weibull distribution, we can state with statistical confidence that the functionalized pool is 1.36 times more likely to find catalysts with a rate of 0.0047/min or better than the nonfunctionalized pool. This is achieved by the mere addition of a positively charged lysine like functional group to DNA.

Qualitatively both a theoretically derived probability function (aexp(-bk)- aexp(-bk)(exp-kt) and the Weibull distribution are able to derive comparable profiles of distributions of catalysts between the endowed DNA libraries and nonendowed DNA libraries. These two distributions clearly show that the initial distribution for the functionalized pool has a greater number of faster catalysts than the nonfunctionalized pool.

A rigorous statistical analysis showed that the average rate between the functionalized pool and the nonfunctionalized pool possessed confidence intervals that did not overlap. This implied that the rates between the functionalized base and nonfunctionalized pools were statistically significant and were unlikely due to stochastic uncertainty. Thus any subsequent transformation of the data, including the aforementioned quantitative statement on functionalized pool improvement would be statistically significant.

The trends and behavior of the selection experiments have been predicted using probability theory and subsequently measured with some degree of statistical significance. We can now predict and measure the rate of optimization by the selection pressure, the consequence of tightening the selection pressure on the population of catalysts, and most importantly, quantify the improvement of finding a selectable behavior or property in one type of biopolymer over another.

The results from the *in vitro* experiments themselves are also of particular significance. First, although the conditions for selection reproduced as closely as possible those used by Breaker and Joyce and the molecules yielded catalytic efficiencies comparable to those obtained by their experiments, the sequences themselves selected

from the natural library were not similar to the sequences obtained by Breaker and Joyce. This suggests that the landscape relating catalytic power to sequence space is sufficiently rugged (Beamer 1988), the distribution of catalytic power across this landscape is sufficiently sparsely, and/or the selection pressure is not sufficiently strong to enforce convergence to the same sequences. This is different from the results obtained by parallel experiments with DNA molecules that act as receptors for ATP (Battersby 1999 Huizenga, 1995) and suggests that catalysts are more difficult to find than receptors.

Also, the sequences of the catalysts containing the J base were not analogous to those extracted from the natural DNA library. This suggested that the J base contributed uniquely to the ability of these molecules to catalyze the cleavage of an RNA-DNA linkage. To confirm this suggestion the catalyst sequences were prepared with T replacing J. No catalytic activity was detectable in these molecules, establishing the essential nature of the J-base for catalysis.

Finally it is worth noting that throughout the selection, including in the modest sample of sequences examined individually, catalysts derived from a library containing the J-base outperformed catalysts derived from the natural DNA library. In one individual instance (J729 Table 3), the J sequence performed 25 times better than the best natural deoxyribozyme. This suggests that the J base increases the intrinsic ability of a library to deliver catalytic power.

Simply put, both the quantitative data and qualitative data show that functionality improves DNA catalysis. This is the first time that data of this type have been obtained for any combinatorial experiment. The results are consequential for any effort to generate a

combinatorial solution to a technological problem and for any discussion that suggests a combinatorial origin for life on Earth.

In closing, this project has introduced an unprecedented way of approaching *in vitro* selection and looking at the origins of life. The methodology and experimental data for the catalytically active functionalized DNA molecule have proven to be equally important. Both concepts give us an idea of how feasible it is to do *in vitro* selection with functionalized bases and at the same time, gain some new insights on the molecular evolutionary aspects on the origins of life. Functionalized DNA moieties may have existed at one time as substrates for Archaeal thermophiles, or they could themselves have been self replicating molecules. The closer we are at obtaining a more complex molecule, the closer we become to a self-sustaining one. The first step taken here makes it worth articulating the project statement. A random pool of amino acid-like DNA molecules has been evolved under Darwinian selection pressures to perform not only catalysis but a quantifiable improvement in the rate of catalysis over non-modified catalytic DNA. As a result a bridge between the chemical system and life has been formed.

REFERENCES

Anders E. (1989) Pre-Biotic Organic Matter from Comets and Asteroids Nature 342, 255 257

Bartel D P., and Szostak Jack W (1993) Isolation of New Ribozymes from a Large Pool of Random Sequences. Science 261, 1411 1418

Battersby T R. Ang D N., Burgstaller P Jurezyk, S Buchanan, D D. Bowser M T Kennedy R.T., and Benner S A (1999) An *In Vitro* Selection for Adenosine Binding Aptamers Using a Fluorionalized Base. Journal of the American Chemical Society In press

Benner S A., and Ellington A D (1988). Interpreting the Behavior of Enzymes. Purpose or Pedigree? CRC Crit Rev Biochem. 23 369-426

Blackburn G M. and Gait M.J (1996) Nucleic Acids in Chemistry and Biology 2nd Edition. New York; Oxford University Press

Bjork, G R (1995) Biosynthesis and Function of Modified Nucleosides. In tRNA Structure Biosynthesis and Function U L R Dieter Soll, ed. Washington ASM Press, 165-207

Blanco L and Margarita, A S., (1995). Mutational Analysis of Bacteriophage I29 DNA Polymerase. Methods in Enzymology 262 283 293

Bock, L. C., Griffin Linda C Latham, John A., Vermaas, Eric H and Toole John J (1992) Selection of Single-Stranded DNA Molecules that Bind and Inhibit Human Thrombin Nature 355 564-566

Braithwaite, J 1 Ito D K. (1991). Compilation and Alignment of DNA Polymerase Sequences Nucleic Acids Research 19 4045-4057

Breaker R R. (1997) In Vitro Selection of Catalytic Polynucleotides Chemical Review 97 371 390

Breaker R R., Joyce G. (1995) A DNA Enzyme with Mg^{2+}-Dependent RNA Phosphoesterase Activity Chemistry and Biology 655 660

Burgstaller P and Famulok M (1994) Isolation of RNA Aptamers for Biological Cofactors by *In Vitro* Selection Angewandte Chemie-International 33, 1084-1087

Burke D H Gold L (1997) RNA Aptamers to the Adenosine Moiety of S adenosyl Methionine Structural Inferences from Variations on a Theme and the Reproducibility of SELEX Nucleic Acids Research 25 2020-2024.

Cate, J H Gooding, A R Podell E, Zhou K Golden B L., Kundrot, C E., Cech T R, and Doudna, J (1996). Crystal Structure of a Group I Ribozyme Domain Principles of RNA Packing. Science 273 1678-1685

Copeland W C., Deng, Q., and Wang T F (1995) Rationale for Mutagenesis of DNA Polymerase Active Sites DNA Polymerase a Methods in Enzymology 262 284-303

Creighton, S Bloom Linda B and Goodman, M F. (1995). Gel Fidelity Assay Measuring Nucleotide Misinsertion Exonucleolytic Proofreading, and Lesion Bypass Efficiencies In Methods in Enzymology Academic Press Inc.) pp 232-257

Dai, X., De Mesmaeker, A., and Joyce, G F (1995) Cleavage of an Amide Bond by a Ribozyme. Science 267 237-240

Doudna, J A., and Cate, J H (1997). RNA Structure, Crystal Clear? Current Opinion in Structural Biology 7 310-316.

Eigen M (1971) Selforganisation of Matter and the Evolution of Biological Macromolecules. Die Naturewissenschaften 58 465 522

Ekland, E. H., and Bartel D P (1996) RNA-Catalysed RNA Polymerization Using Nucleoside Triphosphates Nature 382, 373-376

Ellington A D., and Szostak J W. (1992). Selection *In Vitro* of Single-Stranded DNA Molecules that Fold into Specific Ligand-Binding Structures. Nature 355 850-852.

Esteban J A., Walter N G Kotzorek, G Heckman, J E., and Burke J M (1998) Structural Basis for Heterogeneous Kinetics Reengineering the Hairpin Ribozyme. Proceedings of the National Academy of Sciences 95 6091-6096

Ferre-D'Amare, A. Zhou K. and Doudna, J A (1998). Crystal Structure of a Hepatitis Delta Virus Ribozyme. Nature 395 567-574

Geiger A Burgstaller P Von der Eltz H, Roeder A and Famulok M (1996) RNA Aptamers that Bind L-arginine with Sub-Micromolar Dissociation Constants and High Enantioselectivity Nucleic Acids Research 24 1029 1036.

Gesteland, R. F., Cech, T R. and Atkins J F (1999) The RNA World 2nd Edition Plainview New York Cold Spring Harbor Laboratory Press

Geyer C R Sen D (1997) Evidence for the Metal Cofactor Independence of an RNA Phosphodiester-Cleaving DNA Enzyme Chemistry & Biology 4 579-593

Golden B. L., Gooding A. R., Podell E. R., and Cech T. R. (1998) A Preorganized Active Site in the Crystal Structure of the Tetrahymena Ribozyme. Science 282, 259-264

Green, R., Switzer C., and Noller H. F. (1998) Ribosome-Catalyzed Peptide-Bond Formation with an A-Site Substrate Covalently Linked to 23S Ribosomal RNA. Science 280, 286-289

Hampel A., Cowan J. A. (1997). A Unique Mechanism for RNA Catalysis: The Role of Metal Cofactors in Hairpin Ribozyme Cleavage. Chemistry & Biology 4, 513-517

Hopfner K. P., Eichinger A., Engh R. A., Laue, F., Ankenbauer W., Huber and Angerer, B. (1999) Crystal Structure of a Thermostable Type B DNA Polymerase from Thermococcus gorgonarius. Proceedings of the National Academy of Sciences 96, 3600-3605

Huizenga, D. E. and Szostak, J. W. (1995) A DNA Aptamer That Binds Adenosine and ATP. Biochemistry 34, 656-665

Illangasekare M., Kovalchuke O., Yarus M. (1997) Essential Structures of a Self-Aminoacylating RNA. Journal of Molecular Biology 274, 519-529

Irvine, D., Tuerk, C., and Gold, L. (1991) Selexion. Systematic Evolution of Ligands by Exponential Enrichment with Integrated Optimization by Non-Linear Analysis. The Journal of Molecular Biology 222, 739-761.

Kim, Y., Eom S. H., Wang J. M., Lee D. S., Suh S. W., Steitz T. A. (1995) Crystal-Structure of Thermus Aquaticus DNA Polymerase. Nature 376, 612-616.

Jaeger, L. (1997). The New World of Ribozymes. Current Opinion in Structural Biology 7, 324-335

Johnson, N. L., Kotz, S., and Balakrishnan, N. (1994) Continuous Univariate Distributions, volume 1, 2nd Edition. New York: John Wiley & Sons, Inc.

Joyce, G.F. (1989) RNA Evolution and the Origins of Life. Nature, 338, 217-224

Lazcano, A. and Miller, S. L. (1996). The Origin and Early Evolution of Life: Prebiotic Chemistry, the Pre-RNA World, and Time. Cell 85, 793-798

Lee, D. H., Severin, K., Yokobayashi, Y., and Ghadiri, M. R. (1997) Emergence of Symbiosis in Peptide Self-Replication through a Hypercyclic Network. Nature 390, 591-594

Lee, D. H., Granja, J. R., Martinez, J. A., Severin, K., and Ghadiri, R. (1996) A self-replicating peptide. Nature 382, 525

Lott, W. B., Pontius, B. W., von Hippel, Peter H. (1998). A Two-Metal Ion Mechanism Operates in the Hammerhead Ribozyme-Mediated Cleavage of an RNA Substrate. Proceedings of the National Academy of Sciences 95: 542-547.

Nelson, W. (1982). Applied Life Data Analysis. 1st Edition. New York: John Wiley & Sons, Inc.

Nolte, A., Klußmann, S., Bald, R., Erdmann, Volker A., and Fürste, J. P. (1996). Mirror-Design of L-Oligonucleotide Ligands Binding to L-arginine. Nature Biotechnology 14: 1116-1119.

Oparin, A. I. (1978). Modern Concepts of Origin of Life on Earth. Scientia, 113: 1-4.

Prudent, J. R., and Schultz, P. G. (1995). RNA Catalysis and Transition State Stabilization. Nucleic Acids and Molecular Biology 10: 99-106.

Prusiner, S. B. (1997). Prion Diseases and the BSE Crisis. Science 278: 245-251.

Rich, A. (1962). On the Problems of Evolution and Biochemical Information Transfer. Horizons in Biochemistry: Kasha & Pullman, eds: 103-126.

Robertson, M. P., and Miller, S. L. (1995). An Efficient Prebiotic Synthesis of Cytosine and Uracil. Nature 375: 772-375.

Robertson, M. P., Miller, S. L. (1995). Prebiotic Synthesis of 5-Substituted Uracils: A Bridge Between the RNA World and the DNA Protein World. Science 268: 702-704.

Sakthivel, K., Barbas, C. F. Expanding the potential of DNA for binding and catalysis: Highly functionalized dUTP derivatives that are substrates for thermostable DNA polymerases. (1998). Angew. Chem. Int. Edition 37: 2872-2875.

SantaLucia, J. (1998). A Unified View of Polymer, Dumbbell, and Oligonucleotide DNA Nearest-Neighbor Thermodynamics. Proceedings of the National Academy of Sciences 95: 1460-1465.

Santoro, S. W., and Joyce, G. F. (1998). Mechanism and Utility of an RNA-Cleaving DNA Enzyme. Biochemistry 37: 13330-13342.

Sassanfar, M., Szostak, J. W. (1993). An RNA Motif that Binds ATP. Nature 364: 549-553.

Scott, W. G., Murray, J. B., Arnold, J. R. P., Stoddard, B. L., and Klug, A. (1996). Capturing the Structure of a Catalytic RNA Intermediate: The Hammerhead Ribozyme. Science 274: 2065-2096.

Scott W G Finch J T , and Klug A (1995) The Crystal Structure of an All-RNA Hammerhead Ribozyme A Proposed Mechanism for RNA Catalytic Cleavage. Cell 81 991-1002.

Seelig, B , and Jaschke A (1998) A Small Catalytic RNA Motif with Diels Alderase Activity Chemistry & Biology 6 167 176

Smith D., Kirschenheuter G P Charlton J Guidot D M., and Repine J.E. (1995) *In Vitro* Selection of RNA Based Irreversible Inhibitors of Human Neutrophil Elastase Chemistry & Biology 2. 741-750

Steitz, T A (1998) A Mechanism for all Polymerases. Nature 391 231 232

Stribling R., Miller S L. (1987) Energy Yields for Hydrogen Cyanide and Formaldehyde Synthesis the HCN and Amino Acid Concentrations in the Primitive Ocean. Origin of Life, Evolution, and the Biosphere 17 261-273

Szostak, J W (1998) Redesigning the Molecules of Life Springer-Verlag)

Tarnow T.M., Tarnow S L. Eaton B.E. (1997) RNA-Catalysed Carbon-Carbon Bond Formation Nature 389 54 57

Thompson, J D Higgins, D G and Gibson, T J (1994). CLUSTAL W Improving the Sensitivity of Progressive Multiple Sequence Alignment through Sequence Weighting Position-Specific Gap Penalties and Weight Matrix Choice Nucleic Acids Research 22, 4673-4680.

Unrau P.J., and Bartel D P (1998) RNA-Catalysed Nucleotide Synthesis. Nature 395 260 263

Wackerly D D., Mendenhall W Scheaffer R. L. (1996) Mathematical Statistics with Applications. 5th Edition Belmont CA Wadsworth Publishing Company

Wang J Sattar A. K. M. A., Wang, C. Karam J D Konigsberg, W. H., and Steitz, T A (1997) Cell 89. 1067-1099

Weibull, W (1951) A Statistical Distribution Function of Wide Applicability Journal of Applied Mechanics 293-297

Wilson, C and Szostak, J W (1995) In vitro evolution of a self-alkylating ribozyme Nature 374 777 782.

Zhang, B and Cech T. R (1997) Peptide Bond Formation by In Vitro Selected Ribozymes Nature 390, 96 100

BIOGRAPHICAL SKETCH

Darwin Noel Ang was born in Stillwater, Oklahoma on December 9, 1971. He is the eldest of two sons of Professor James S. Ang and Mrs. Josephine G. Ang. In the fall of 1990, he attended the Florida State University where he received his B.S. in chemistry and anthropology. Originally he aspired to be an archaeologist or writer, but found his passion in helping others especially through research and medicine. In 1993 he received FSU's President's Humanitarian award.

In 1994, Darwin attended the University of Florida where he was accepted into the M.D./Ph.D. Medical Scientist Training Program. After completing two years of medical school, he began his graduate training under the guidance of Professor Steven A. Benner. During his graduate studies, he was partially funded by the American Heart Association Predoctoral Fellowship.

After receiving his Ph.D., Darwin plans to return to medical school to complete his clinical training and graduate with the class of 2001. In the meantime, he enjoys playing on his piano and playing men's club soccer with his friends where he serves as captain and player-coach.

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy

Steven A. Benner, Chair
Professor Anatomy and Cell Biology

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy

Benjamin A. Horenstein
Associate Professor of Chemistry

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy

Alfred A. Lewin
Professor of Molecular Genetics and Microbiology

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy

Gerry F. J. Shaw
Professor of Neuroscience

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality as a dissertation for the degree of Doctor of Philosophy

Stephen F. Sprague
Associate Professor of Anatomy and Cell Biology

This dissertation was submitted to the Graduate Faculty of the College of Medicine and to the Graduate School and was accepted as partial fulfillment of the requirements for the degree of Doctor of Philosophy

August 1999

Dean, College of Medicine

Dean, Graduate School